MAN-MACHINE EFFECTIVENESS ANALYSIS

A SYMPOSIUM OF THE HUMAN FACTORS SOCIETY - LOS ANGELES CHAPTER

Edited By

Robert E. Blanchard
Chief Scientist
Western Division,
Dunlap and Associates, Inc.

Douglas H. Harris
Group Scientist
Autonetics,
A Division of North American Aviation, Inc.

Papers Presented at a Symposium Sponsored
by the Human Factors Society - Los Angeles
Chapter at the University of California at
Los Angeles, 15 June 1967

AN-MACHINE EFFECTIVENESS ANALYSIS

SYMPOSIUM OF THE HUMAN FACTORS SOCIETY - LOS ANGELES CHAPTER

CONTENTS

PREFACE

This book is a result of the fourth annual symposium sponsored by the Human Factors Society , Los Angeles Chapter to promote the exchange of information among behavioral scientists concerned with man-machine systems. This symposium, "Man-Machine Effectiveness Analysis: Techniques and Data Requirements, " was conducted on 15 June 1967 at the University of California at Los Angeles. Robert Blanchard was General Chairman; he was assisted by Douglas Harris, Meredith Mitchell, Jack Parrish, Russell Smith, John Stroessier, Alan Swain and Wilson Wong.

The support and cooperation provided by the University of California at Los Angeles and Autonetics, A Division of North American Aviation, are gratefully acknowledged. UCLA provided the facilities for the symposium and Autonetics prepared the final layout of this book.

R.E.B.

D.H.H.

1. THE CHALLENGE

The increased cost and complexity of modern man-machine systems have directed attention toward methods for predicting and evaluating system effectiveness. As a result, a new technology, generally referred to as system effectiveness analysis, has developed. The essential emphasis of this technology is the identification and quantification of critical design factors, and the development of models which relate these factors to system effectiveness.

Since human performance is critical to the effectiveness of most man-machine systems, techniques for dealing with human factors are needed. However, while notable progress has been made in handling the machine aspects of systems, only limited attention has been directed toward the development of techniques for quantifying human performance and relating human factors to system effectiveness. The first major attempt to organize and present the thinking of individuals engaged in research relevant to this problem was a symposium/workshop held in New Mexico in 1964. It was sponsored jointly by the Human Factors Subcommittee of the Electronics Industries Association and the University of New Mexico. Selected papers from the symposium were published in Human Factors. In 1966, a session of the American Psychological Association was devoted to the reliability of human performance; three papers were presented. Then, in January of this year, the Navy Material Command and the National Academy of Engineering sponsored a symposium on the subject of human performance quantification in system effectiveness in Washington D. C. Although some other technical meetings have dealt with related areas, symposia directed toward the central problem of dealing with human performance in man-machine effectiveness analysis have been limited to these three. In planning the meeting that resulted in this book, it was our feeling that those symposia could be complemented by one which directed its attention to recent developments in models, data and techniques.

Interest in the man aspects of system effectiveness analysis appears to be growing; behavioral scientists are being challenged to provide the required models, data and techniques. There are some general man-machine modeling techniques currently available such as Technique for Establishing Personnel Performance Standards,[1] Technique for Human Error Rate Prediction[2] and Operator Overload Prediction Technique.[3]

---

[1] Mitchell, M. B., Smith, R. L., & Verdi, A. P. Development of a Technique for Establishing Personnel Performance Standards (TEPPS). Phase III - Final Report, Dunlap and Associates, Inc., Santa Monica, July 1966.

[2] Swain, A. D. A Method for Performing a Human Factors Reliability Analysis. Sandia Corporation Monograph SCR-685, Albuquerque, N. M., August 1963.

[3] Siegel, A. I., & Wolf, J. J. A Technique for Evaluating Man-Machine System Design. Human Factors, 1961, 3, 18-28.

Other researchers have been pursuing the problem from a system-specific point of view; their findings may eventually contribute to more general approaches. Even so, we appear to be at an elementary stage of development. In his concise, critical review of presently available approaches to the quantification and prediction of man-machine operability, Freitag[4,5] concluded that ".... a practical procedure having the required validity and reliability for establishing contractual operability minima appears to be some years away."

Equally important is the consideration of the types of data required. All models developed to date require some form of data on the human activities required by modern, complex systems. Most available data are point estimates gleaned from the experimental literature. The Data Store prepared by the American Institutes for Research[6] contains data that is limited in behavioral description and is questionable in validity due to the necessity of extrapolating from laboratory studies to field situations. Some work is underway to develop human performance data banks within companies or military activities to meet specific needs; however, these data are not as yet generally available. Some interim procedures [7,8,9] using scaling techniques have been employed to obtain estimates of human performance values. Since it is apt to be some time before a generally applicable, available store of human performance data exists, it is apparent that some interim reliance will be placed on these techniques if human factors are to receive consideration in system effectiveness studies.

[4]Freitag, M. Quantification of Equipment Operability: I. Review of the Recent Literature and Recommendations for Future Work. U.S. Navy Electronics Laboratory Memorandum 940, June 1966.

[5]Freitag, M. Review of Quantitative Operability Prediction Techniques: Phase I Final Report. Jakus Associates, San Diego, California, February, 1967.

[6]Munger, S J., Smith, R. W., & Payne, D. An Index of Electronic Equipment Operability: Data Store. Pittsburgh, Pennsylvania: American Institutes for Research Report No. AIR-C43-1/62RP(I). January 1962.

[7]Williams, H. L. Reliability Evaluation of the Human Component in a Man-Machine System. Electrical Manufacturing, 1958, 61, 78-82

[8]Irwin, I. A., Levitz, J. J., and Freed, A. M. Human Reliability in the Performance of Maintenance. Proceedings of the Symposium on the Quantification of Human Performance. Albuquerque, New Mexico: University of New Mexico, 1964, 143-198.

[9]Blanchard, R. E., Mitchell, M. B., & Smith, R. L. Likelihood-of-Accomplishment Scale for a Sample of Man-Machine Activities. Dunlap and Associates, Inc., Santa Monica, June 1966.

In our opinion, there is an urgent need to advance the state-of-the-art in man-machine effectiveness analysis. The challenge to the behavioral science community is one of joining and contributing to the multi-disciplinary effort directed toward the development of more practical and sophisticated analysis approaches. To this end, the recent thinking of 10 behavioral scientists who have been concerned for some time with man-machine effectiveness analysis is presented in this collection of seven papers. Since these 10 scientists were located in seven different organizations and had been employed on a variety of different projects, it should come as no surprise that their points of departure differ. With respect to objectives, however, we find them in agreement, and this is what ties the seven papers together.

Consistent with the basic elements of man-machine effectiveness analysis, the papers are organized into the following three sections — Models, Data and Techniques. In the first section (Models), the problem of allocating system effectiveness requirements among the functional units or states of a system is discussed by Mitchell and Blanchard. Also, in this section, models for dealing with human performance in man-machine effectiveness analysis are discussed in separate papers by Williams and by Mason and Rigney. The second section (Data) consists of papers by Rigby and Meister which discuss obtaining and using data in the quantification and prediction of human performance. In the third section (Techniques), an application of man-machine simulation is presented by Spencer, and a technique for man-machine evaluation is described by Sheldon and Zagorski.

Hopefully, this collection of papers will be useful to those who are confronted with the problem of man-machine effectiveness analysis and those who are working toward the development of better models, data and techniques for handling the problem.

2. THE ALLOCATION OF SYSTEM EFFECTIVENESS REQUIREMENTS FOR MAN-MACHINE EFFECTIVENESS ANALYSIS

Meredith B. Mitchell and Robert E. Blanchard
Dunlap and Associates, Inc.

Allocation of system effectiveness requirements, in its broadest sense, is something we all do all the time. Each one of us strives toward particular goals which at some level of consciousness are considered to possess certain preconceived minimal characteristics. At least the initial steps in approaches we use to achieving those goals -- if we behave rationally -- are somehow evaluated against alternative procedures on the basis of such criteria as (1) the likelihood with which each may be expected to lead to success, (2) the time they require, and (3) their relative emotional, physical, and monetary costs. Each step in an approach is weighed on the basis of its contribution toward achieving the ultimate goal. Presumably, then, we act under the assumption that the sum total of the contributions of the individual steps is at least the very minimum we would expect and desire when the goal is reached.

Of course, human goals are generally likely to be in a state of change or modification as new information and experience add to the store of action determinants. But how many of us consciously define our goals at any given moment in time, sufficiently objectively to be able to specify ahead of time the precise nature of our minimal final requirements? And how often do we perceive, plot and weigh all relevant alternative courses of action to determine if and how we realistically can allocate those requirements, and then test the model so as to select the one which is optimal?

If man's development had emphasized such rigid planning procedures, life would be mechanical, frequently inappropriate and sorely lacking spontaneity, but man-machine effectiveness analysis would certainly be much easier. As it is, we find ourselves faced with man's propensity (1) to define his objectives in rather vague terms and (2) to define his requirements not at all. Perhaps it is because of this limited past experiences that he forms a narrow repertory of approaches to problems and develops a tendency to move in relation to a goal with a trial-and-error or familiar-but-not-necessarily-optimal set of motions. Thus, for effective effectiveness analysis and allocation, we must overcome awareness of both uncertainty and possibly of change in order to be able to objectify without closing the door to heuristics.

To develop a method for allocating effectiveness requirements needs three basic questions to be answered; interestingly, once the method exists, it must be able to answer the same three questions:

1. Allocation of what?
2. Allocation to what?
3. Allocation with what?

    a. Tools (rules)?
    b. Material (input data)?

In a general way, this paper is addressed to answering those questions from the point of view of methodological development. The primary goal, however, will be to indicate some of the problems which arise in attempting to allocate system effectiveness requirements (SERs) in man-machine models.

Characteristics of System Effectiveness Requirements -- That Which is to be Allocated

Effectiveness is generally defined as the degree to which the system (or a functional unit of a system) is able to achieve its stated objective. Quantification of effectiveness requires the identification of one or more measurement dimensions. The most frequently used dimensions are accuracy, time, quantity and rate, constrained by cost limitations. Effectiveness dimensions must be related directly (or as directly as possible) to stated system objectives. In some cases, a composite of effectiveness dimensions may be necessary in order to reflect the system objective adequately.

In order to define an acceptable level of performance with respect to system objectives, a stipulated value or magnitude is established on the performance dimensions, that value constitutes the system effectiveness requirement. For example, an effectiveness dimension of detection range might be selected for a surveillance system; a value of 100 miles, with an expectancy level of .95, might then be determined through mission analysis studies as the system effectiveness requirement. Effectiveness requirements also may be stipulated for major functions the system is to perform. For the example above, effectiveness requirements may be stipulated for such major system functions as target identification, classification and threat assessment.

Effectiveness requirements may take the form of a single value on an effectiveness dimension, or under certain circumstances, several values or an interval may be defined representing levels of effectiveness which are acceptable under specified operating or environmental conditions for that system. In many instances, the system effectiveness requirement is stated as the required probability of achieving a particular level of performance on the dimension, e.g., probability of achieving the required output state at a particular accuracy, time or rate. When more than one effectiveness dimension is necessary in order to reflect the system objective adequately, the effectiveness requirement may be represented as an index resulting from the mathematical combination of values on several effectiveness dimensions.

For allocation, therefore, it is necessary that mission analyses have previously been directed toward defining requirements appropriate for effectiveness analyses. Values along all relevant dimensions must emerge as an end product. In past and even current practices, such end products are sorely lacking, reflecting the haphazard or untested intuitive approach to design for meeting imprecisely defined system objectives. It is rare that effectiveness requirements for a system are specified, either because they had not been considered or because customers do not wish to be faced with the fact that serious objectives may not always be reached -- or because systems analysts are unwilling to record fallibility for all to see.

Allocation to What?

A system's mission-specified objective defines the ideal end product, end result or output state. The effectiveness requirement of a system relates to that objective, conditional upon a definable input state. For example, if

the SER of an assembly line production is that x or more operable articles
be produced per day, it is assumed that all required facilities, parts, equip-
ment, funds, personnel, etc. are available at the outset. Thus, the system
may be considered to be a complex Personnel-Equipment Functional Unit
(PEF Unit) having a definable input state and a mission-required output
state defined by the SER.

The concept of requirements allocation implies a multiplicity of con-
tributors to the meeting of those requirements. In practice, "contributors"
generally have been found to fall into one of two categories of verbal
description: Activities or system states. Some effectiveness analysts
believe that approaches employing description of the operations involved in
PEF Units are as easy to use and yield the same results as approaches
which emphasize system states resulting from activity transitions. That
may sometimes be true. However, it has been our experience that individuals
who are activity-oriented tend to be more stimulus-bound and less free from
pre-conceived notions than those who are state oriented. Figures 2-1 and 2-2
illustrate a generalized hypothetical communication system which will be
used as an example to demonstrate how that tendency seems to arise.

In the example, the system PEF Unit activity can be defined as
"transmit data a, b and c from A to B, " given that all conditions are "GO"
for A to contain those data and for potential communication between A and B.
Equivalently, one could specify only the output state, "B possesses data a,
b and c" given the same "GO" conditions. By logical deduction, based upon
currently conceivable communication systems, two intermediate states
(or three lower-level PEF Units) can be defined; those are identified as
States I and II in Figure 2-2.

To imply those states by describing the activities of PEF Units #1 and
#2 one may inadvertently restrict thinking to particular modes of operation.
Statements like "A establishes contact with B . . ." or "A and B confirm
each other's identity . . ." tends to imply a verbal communication between
two persons. However, the system may be in the design stage when it
would be desirable to consider alternatives and perform tradeoff analyses.
It might be consistent with the system's objectives and requirements to
consider possible hardline communications between two computer systems
or between a human and a remotely controlled vehicle or between a signalling
satellite and ground station. In contrast, specification of system states tends
less to imply transitionary methods for achieving those states; rather, there
are many possible and feasible methods, consideration of which depends
upon the experience and creativity of the analyst.

Emphasis on system states also guides the analyst to clear and concise
specification of required input states. In the example, State II not only
requires that the data be available for transmission, but also that (1) the
data are needed at the receiving end and (2) there are measurable criteria
for ascertaining that appropriate and errorless contact is made before
transmission. For some reason, activity-oriented people often fail to per-
ceive input state requirements with clarity.

Specification, or consideration of required system states tends to
lead to a creative, open-minded approach to analysis -- both for new designs
and for evaluation of existing systems. For new designs, one remains open
to alternative approaches to satisfying requirements; for existing systems,
one may look at existing procedures, seek the requirements they are intended
to meet, then ascertain if (1) the procedures actually do meet the requirements,
and (2) if the requirements could possibly be met in some other, unspecified,
and more or less effective manner.

INPUT STATE:
DATA A, B AND C EXISTS AT A
AND CAPABLE TO TRANSMIT;

CHANNEL BETWEEN A AND B
CAPABLE OF BEING APPLIED;
B CAPABLE OF RECEIVING

TRANSMIT DATA
a, b AND c
FROM A TO B

OUTPUT STATE:
B POSSESSES

DATA a, b AND c
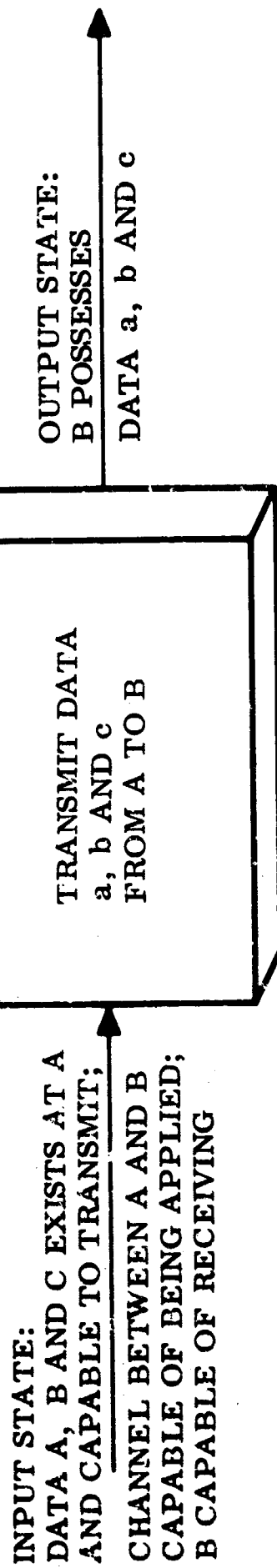
Figure 2-1. Generalized Hypothetical Communication System

INPUT STATE:

DATA a, b AND c
EXISTS AT A AND
CAPABLE TO TRANS-
MIT; CHANNEL
BETWEEN A AND B
CAPABLE OF BEING
APPLIED; B CAPABLE
OF RECEIVING

CRITERIA FOR
COMMUNICATION

#1

STATE I

A
ESTABLISHES
NEED FOR
CONTACT
WITH B

DATA a, b AND
c ARE AVILABLE
A' A; CRITERIA
FOR COMMUNICA-
TION ARE MET

B REQUIRES
DATA a, b AND
c FROM A

#2

STATE II

A ESTABLISHES
CONTACT WITH
B

CRITERIA FOR
ASCERTAINING
ACCURACY OF
CONTACT

COMMUNICATION
CHANNEL IS OPEN
BETWEEN A AND B
AND NOT BETWEEN
A NOR B AND ANY
OTHER TERMINAL

#3

A COMMUNICATES
DATA TO B

OUTPUT STATE:
B POSSESSES
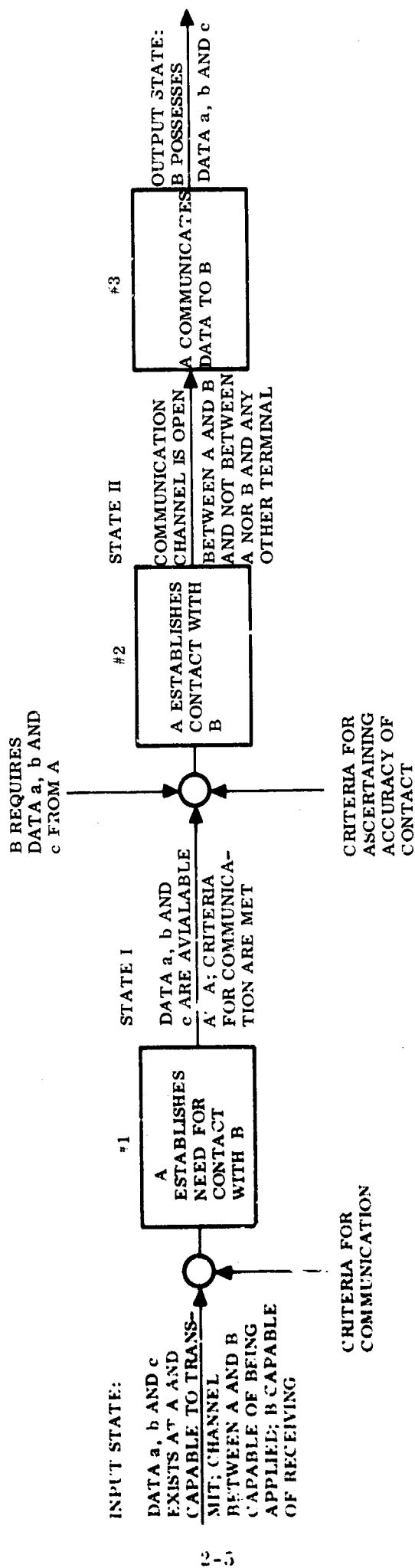DATA a, b AND c

Figure 2-2. First Level Partitioning of Figure 2-1.

Finally, the specification of intermediate system states focusses attention on within-systems effectiveness requirements. Ideally, it is to these that allocation would be directed. However, particularly at more detailed levels of specificity, each state is conditional upon prior states unique for the particular system under consideration. As a result, the data necessary for allocating to states tend to be system-related. This is probably the single greatest drawback to state-based analyses, sets of input and output states for any one system tends to apply uniquely to that system.

In contrast, activities can be sufficiently segmented so that verbal descriptions appear to be generalizable across systems. As a result, existing data stores are activity-oriented. Their primary disadvantages, however, are that they are unrelatable to system contexts and that they do not combine in a simple manner.

To make problems more difficult, certain types of man-machine activities do not lend themselves to either state or activity analysis at a detailed level of specificity. Two obvious kinds of such operations are complexly contingent tracking-type and decisions-making tasks. When the rules for such activities as tracking or decision-making are difficult to verbalize and depend largely on intuition developed from extensive experience, the analyst is hard-pressed to do better than consider the contribution to system effectiveness of the total complex activity.

Thus, allocation of effectiveness requirements for a large system may need to be directed simultaneously toward both simply defined and complex, critical transitions (or the states defining those transitions). At present, there appears to be a need for a method of combining the activity and state approaches to develop generally-applicable, reliable man-machine units of performance and for generating equivalently useful units for all types of activities -- if that is possible.

Bases for Allocation

It is necessary but not sufficient that valid system effectiveness requirements exist and are derived from mission analyses, and that the system is partitioned into manageable units for evaluation of their contribution to system performance. There still remains the need for relevant, internally consistent data and procedural rules for systematically applying those data to enable allocation of a given system's SERs among its component units.

Whether a state or activity approach is used, it would be ideal if there were some bases upon which allocation could be performed at progressively more specific levels of verbal description. In Figure 2-2, analysis would be greatly simplified if it were clear that each state-to-state transition always contributed the same relative amount to the success of the system, independent of the means by which the transition is implemented. For example, assume that (1) concern is with the probability (P_o) of successfully achieving the output state, (2) each PEF Unit's output state is independently conditional upon its total input state, (3) the existence of a PEF Unit's output state implies its input state, and (4) each transitional PEF Unit is somehow known to contribute equally to system success. Under those assumptions, the conditional probability of each output state given its input state would be $(P_o)^{1/3}$. It would then be possible to treat each PEF Unit as a complete system, generate approaches to meeting the requirements of its output state and generate (in a creative way) progressively more specific and alternative means for achieving those states.

The procedure in the above example implies the existence of data which indicates that the contribution of the three PEF Units are equivalent, independent of the means by which they are performed. While the results would limit consideration of possible intermediate states and methods of achieving those states, the most serious problem is evaluating the validity of the equivalence rule in the first place. The procedure also draws on probability theory for its multiplicative rule relating to independent events: in the example, the events were considered independently conditional.

But problems arise when it becomes evident that some system transitions are more or less dependent upon others (i.e., when certain states are distributed along a kind of feedback dimension to alter the distribution of prior state dimensions). Both the magnitude and target(s) of dependencies are frequently difficult to define. We need techniques for defining and handling degrees of dependency.

Furthermore, even if all transitions were independent, there would still be the problem of relating to the overall SER the distributions along the effectiveness dimensions of each system state. If all states could be dichotomized (go/no-go) such that the dichotomies applied each time the system were exercised, the problem would be immensely simplified. In other words allocation could be used to specify the cut-off point separating success from failure. Often, however, cut-off points vary along on effectiveness dimension.

Thus, as was indicated earlier, there appears to be a need not only for man-machine performance data, but for multi-dimensional distributions of those data -- such as a level of confidence in successful performance as a function of (1) accuracy, (2) performance time, (3) equipment (reliability and maintainability) needs and costs, (4) personnel training and selection costs, (5) and backup (e.g., operational redundancies and man/equipment logistics). And these data need to be formatted so as to enable relating them to overall effectiveness requirements of the system. Such formatting depends on an allocation procedure which is sufficiently advanced to anticipate application of not-yet-existing data.

Summary and Conclusions

The allocation problem in man-machine effectiveness analysis concerns the accurate determination and specification of the effectiveness requirement of a system, and the development and application of a set of rules by which the system effectiveness requirement, in its various forms, can be distributed among the man-machine functional units/states comprising the system. The resultant allocation must provide a set of performance requirements or standards at a level sufficiently elemental to facilitate (1) trade-off studies, (2) relative appraisal of various system design concepts, and (3) absolute evaluation of a given design concepts, and (4) absolute evaluation of a given design against the system effectiveness requirements established for the system.

To develop a procedure for effectiveness allocation, guidelines must be generated for (1) specifying the system effectiveness requirement along all its dimensions, (2) partitioning the system into meaningful and useful segments and states, (3) characterizing and specifying input data, and (4) relating the SER to system segments consistent with the input data. While current techniques necessarily involve poorly specified requirements, limited or estimated data, and relatively simple rules, the results have been rewarding. At the very least, attention has been turned toward the need for objectifying goals.

Hopefully, more and more complex relations among goals and the steps leading to those goals will be examined sufficiently systematically to enable accurate allocation of effectiveness requirements "resources" in the future.

3. DEPENDENT MODELS FOR ESTIMATING HUMAN PERFORMANCE RELIABILITY

Herman L. Williams
Martin Marietta Corporation

Tasks performed by human operators, maintenance technicians, and ground crews in assembly, test, and handling frequently have a significant effect on the efficiency of a weapon system. An error made by an operator in setting a dial, operating a control, or reading a meter can result in loss of life as well as destruction of equipment worth millions of dollars. Failure of a maintenance technician to diagnose a malfunction or meet a schedule for repair of a component can seriously affect the availability of equipment. Mistakes in assembly, test. and handling can lead to an aborted mission or delivery of an ineffective weapon.

Because of the importance of people in a weapon system, there is an urgent need not only to assign functions properly and to design equipment for ease of operation but to assess the ability of system personnel to perform their assigned tasks. One well-known approach for establishing design feasibility is to construct a time-line and determine if the tasks can be performed in the available time. This approach, although essential, does not complete the evaluation. A man can fail in the performance of a task, even though adequate time is available. In assessing system and design feasibility, therefore, one must also determine the reliability of human performance.

Methods have been developed for estimating human performance reliability. These require that the operational, maintenance, or handling task be broken down into discrete steps. A probability model, which takes into account the arrangement of task steps as well as the relationship of steps to one another, is then fitted to the task. Values are estimated for each element in the model. The probability of success is then computed for the total task.

If discrete steps in a task are independent, one can estimate human performance reliability without undue difficulty. Unfortunately, if steps in a task are performed by a single operator or by operators working together, a dependent relationship occurs, causing much difficulty for the analyst attempting to assess human performance reliability. Models for taking the dependent relationships into account are composed primarily of conditional probabilities arranged mathematically to represent steps in a set of operating procedures. The value of the conditional probability for a given step depends not only upon the immediate circumstances under which the step is performed (i e., equipment design features, environment, etc.) but also upon the particular combination and characteristics of task steps preceding it in the sequence of operations. Sources of probability data available for estimating such values can take the immediate circumstances into account. Unfortunately, the combination of characteristics of earlier steps in a task usually are unique, and the analyst, in attempting to estimate the conditional probabilities, finds that neither data nor procedures are available to help him take the dependent relationships into account.

Background of the Problem

To establish a basis for analyzing the problem, it is necessary first to examine the requirement for and the approach used to obtain estimates of human performance reliability. Such estimates are needed during the concept, design and development, and utilization stages of a weapon system. The concept stage is a period during which a number of alternatives are evaluated to determine which best meets the system objectives and constraints. Reliability of human performance is an important parameter in these evaluations. To be feasible, a system concept must show an acceptable level of reliability for the human operator; therefore, in selecting the best of several feasible concepts, the analyst should consider human performance reliability as one of the major system parameters to be optimized.

During the concept stage, actual equipment and personnel are seldom available for purposes of testing. Comparisons of alternatives are made primarily by means of paper-and-pencil analyses. Steps performed in these analyses are as follows:[1]

1. Definition of mission requirements, which includes the identification of mission objectives, determination of anticipated use environments and mission success criteria, and specification of any other information defining the use conditions of the system.

2. Determination and description of tentative system and equipment design features for each concept, the primary objective of which is to establish the characteristics of the operator-equipment interface. Since the interface includes both operators and equipment, the system description likewise must cover both.

3. Preparation of hypothetical operating procedures, arranged as discrete steps of operator tasks that form the basis for elements in the probability models. Therefore, in preparing hypothetical operating procedures for the system concepts being evaluated, one lists procedural steps, along with sufficient descriptive information to permit probability-of-success estimates to be made.

4. Construction of probability models, which starts with construction of models for subtasks. The outputs from these models are then combined into models representing several subtasks. Outputs from the combined models are in turn combined at progressively higher levels until a model is obtained that represents performance of the total task.

5. Estimation of values for terms in probability models. The approach for estimating probability values for independent events differs considerably from that for estimating dependent probabilities. If the terms in the model are independent, one may estimate the value for a given term without concern for other steps in the operating procedure. In contrast, if the terms are dependent, one must consider earlier steps in the procedure when estimating the value of a given probability.

[1] Williams, H. L., Human Performance Reliability in Operational and Maintenance Tasks. OR 8729. Martin Marietta Corporation, Orlando, Florida. January 1967.

6. Computation of human performance reliability, which proceeds in accordance with the mathematical relationships set forth in the probability models.

Steps in the procedure for estimating human performance reliability during design and development are essentially the same as those used during the concept stage. The concept is fixed by the time the system enters design and development. Alternatives to be considered and evaluated now are limited to system and equipment design features. Human performance reliability is one of the measures used for comparing alternatives and arriving at an optimum design.

During the utilization stage, estimates of human performance reliability are needed for mission and logistics planning purposes. Design features of the system and equipment are no longer tentative. Operating procedures are firm. If adequate test and field data are available, the step-type procedure outlined above is not used. One obtains the necessary estimates from the test and field data by taking the ratio of operator successes to total number of tests or trials. If adequate data is lacking, however, human performance reliability must be computed, using essentially the same procedure as that used in the concept and design and development stages.

When an analysis is conducted based on the six-step approach outlined above, little difficulty is encountered in the first four steps. Established procedures can be used to define the mission requirements, to determine and describe tentative system and equipment design features, and to prepare hypothetical operating procedures. During construction of the models in step 4, the majority of operating procedures can be represented by series and parallel probability models or by minor modifications of these models. If task steps are independent, the general series model is defined by equation 1.

$$P_S = P(X_1 = 1) \ P(X_2 = 1) \dots P(X_n = 1) \tag{1}$$

P_S is the probability of successful task performance. The X_i, $i = 1$, $2, \dots, n$, represent steps in the series task. The relationship, $X_i = 1$, denotes success in performing step i. Although not used in equation 1, $X_i = 0$ denotes failure in performing step i.

Equation 2 gives the general series model for dependent events. It will be noted that the form of the dependent model is similar to that for independent events.

$$P_S = P(X_1 = 1) \ P(X_2 = 1 \mid X_1 = 1) \ P(X_3 = 1 \mid X_1 = 1, X_2 = 1) \dots$$

$$P(X_n = 1 \mid X_1 = 1, X_2 = 1, \dots X_{n-1} = 1) \tag{2}$$

The first term on the right-hand side of equation 2 is the marginal probability of successful performance of step 1. All other terms in equation 2 are conditional probabilities. Note, however, that there is a term-for-term correspondence between equations 1 and 2. In other words, the form of the models is the same; only the values of the individual terms in the model have changed in going from equation 1 to equation 2

Equation 3 gives the general probability model for independent events in parallel.

$$P_S = 1 - P(X_1 = 0)\ P(X_2 = 0)\ \ldots\ P(X_n = 0) \tag{3}$$

In the parallel model, success in a single step gives successful task performance.

Equation 4 gives the general probability model for dependent events in parallel.

$$P_S = 1 - P(X_1 = 0)\ P(X_2 = 0 \mid X_1 = 0)\ \ldots\ldots\ P(X_n = 0 \mid X_1 = 0,$$

$$X_2 = 0,\ \ldots,\ X_{n-1} = 0) \tag{4}$$

Again, the term-for-term correspondence may be observed as one compares equations 3 and 4. As in equation 2, only the values of individual terms in the model have changed in going from the independent to the dependent model.

The problem of concern in this paper arrives when one reaches step 5 in the computational procedure outlined earlier. Data stores are available for use in estimating values for terms in the probability models for independent events, but not for dependent events. One finds, when analyzing human performance reliability, that the great majority of operational procedures encountered are dependent. One must therefore evaluate the effect of the dependent relationships when estimating values for terms in the models. Unfortunately, data and techniques are not presently available for doing the job.

The problem of estimating values for elements of dependent probability models can be solved only by providing the data and/or techniques needed for taking the dependent relationships into account. In deriving the necessary data and techniques, however, one must consider the anticipated characteristics of future data stores, identification of factors responsible for 'ne dependent relationship, and magnitude of the effect upon probability of successful performance of given steps in an operational task.

Characteristics of Future Data Stores

A probability data store is a tabulation of values representing the probability of successful performance of a defined task or task element by an operator of specified characteristics. Although presently available data stores are limited in the categories of tasks and task elements, environmental conditions, and defined operator characteristics covered, it is not unreasonable to expect that future data stores will cover an extensive range of such categories. It is also possible that the data store will provide distributions of probability values as well as the average or expected values. However, to be economically feasible, the data store must be applicable to a wide range of operations. Probability values listed in the data store must be relevant to common elements of a great variety of systems. The common elements are the individual steps or

operations in a task. In the data store developed by the American Institute for Research[2], for example, the common elements are inputs to the operator, mediating processes, and outputs from the operator for specified task steps. Values in the data store are immediately relevant to terms in the probability model, if task steps or elements are independent. In other words, the values are marginal probabilities. They do not take into account dependent relationships, for to do so would limit the range of tasks to which the data store is applicable. One must conclude, therefore, that the conditional probabilities of a model composed of dependent events will not be found in a data store.

It is evident that a problem confronts the analyst attempting to estimate the conditional probabilities of dependent models. He must make the estimates prior to the time prototype equipment is available for experimental study. Yet, he has no available source of fully relevant data. The problem can only be solved by development of models for making the transition from the marginal probabilities of the data store to the conditional probabilities of the dependent model.

Factors Responsible for the Dependent Relationship

The factors responsible for the dependent relationships of a given task step with earlier steps may be defined as those which have a measurable effect upon the probability of successful performance of the given task step. All the factors exerting such an effect have not been identified. Some, however, are known, although the nature and extent of their effect have by no means been established.

It is well known that design features of equipment operated and/or observed early in a sequence of task steps can affect performance in later steps. Studies of aircraft cockpit instrumentation, for example, have shown that the design of instruments with pointers positioned in the same direction during normal operation facilitates the instrument reading task. The design of controls used in a sequence of operations to make the motions consistent from one operation to another likewise facilitates the control task. Conversely, controls and displays which have conflicting design features will degrade performance.

Although certain design features can affect performance in later steps of an operational task, there has been little systematic study in this area to identify such features. No one to date, for example, has compiled a list of the equipment design features suspected of having an effect on probability of successful performance of subsequent steps. Certainly, before one can construct a transition model for taking into account design features of equipment operated earlier, one must determine the design features responsible for the dependent relationship.

The type of activity required of an operator in one step can also influence his performance in a following step, particularly if the earlier step affects the operator's physical condition. For example, if an operator's vision is adapted to the light level external to an aircraft during search for a target, he may have difficulty adjusting in a subsequent step to the light output from displays in the cockpit. A step which exhausts an

[2] Payne, D. and Altman, J. W., An Index of Electronic Equipment Operability, AIR-C-43-1/62, FR, American Institute for Research, January 1962.

operator will, of course, degrade his performance in subsequent steps. An operator's performance in a monitoring task is affected by the level of activity; he can have too much or too little to do.

Unfortunately, as in the case of equipment design features discussed above, no systematic attempt has been made to identify the operator activities which can affect performance later in a task.

Little is known about the effect of the number of steps in a task upon operator performance. If environmental conditions are unpleasant or if time constraints or other stress-producing factors are present, there may be an interaction effect which improves or degrades performance. Study is needed to determine if such an effect actually exists.

Numerous other factors may be responsible for a dependent relationship among task steps. Such factors include task performance time, elapsed time between task steps, arrangement of task steps in a procedure, etc. Interaction effects among many of the factors may also exist. Certainly, the identification of these factors and the determination of relevant interactions constitute a much needed study program.

Form of the Transition Model

Although much preparatory work remains to be done before actual transition models can be constructed, one can determine the general form of the models by using the techniques of experimental design and analysis[3]. The conditions relevant to a given step in an operating procedure may be considered as independent variables of a linear model. For the purposes of this analysis, the given step will be referred to as the reference step, and it is the step for which a probability value is being sought. The response of interest or output from the linear transition model is the conditional probability value. One can arrange the conditions or independent variables in an n-dimensional matrix, so that the independent variables giving the response represented by the pertinent marginal probability in the data store are all included in cell 1 of the matrix. Other cells in the matrix represent independent variables forming the basis for the dependency relationships with earlier steps.

To illustrate the approach, assume that the only factors affecting the probability of success in the reference step are equipment design features, type of activity performed, and the presence or absence of a time constraint on the task. Table 3-1 gives the number of levels and combinations of the independent variables.

The cell in the upper left-hand corner of table 3-1 is designated as cell P_{11}. It gives the probability of successful performance of step j, the reference step, when the equipment design features and operator activities in the reference step are employed in combination with no time constraint; i.e., A_0, D_0, C_0. With only these conditions present, one can estimate the probability of success in the reference step, P_{11}, by means of a marginal probability value from the data store. Suppose that

[3]W. Mendenhall. An Introduction to Linear Models and the Design and Analysis of Experiments. Manuscript. University of Florida. 1966.

Table 3-1: Matrix of Independent Variables

		D_0	D_1	D_2
A_0	C_0	P_{11}	P_{12}	P_{13}
	C_1	P_{21}	P_{22}	P_{23}
A_1	C_0	P_{31}	P_{23}	P_{33}
	C_1	P_{41}	P_{24}	P_{34}
A_2	C_0	P_{51}	P_{25}	P_{35}
	C_1	P_{61}	P_{26}	P_{36}

the conditions represented by the cell P_{24}, i.e., A_1, D_1, C_1, are present. This would indicate that equipment design feature D_1 is operated in an earlier step, that activity A_1 is performed in an earlier step, and that the task is performed under a time constraint. The effect of A_1, D_1 and C_1 is a change in the probability value from P_{11} to P_{24}. Other cells in the matrix may be interpreted in like manner.

D_0 — design features of equipment operated in reference step

D_1, D_2 = design features of equipment operated in earlier steps

A_0 — activity performed in reference step

A_1, A_2 — activities performed in earlier steps

C_0 — no time constraint

C_1 — time constraint imposed on task

P_{ik} — probability of successful performance of step j. the reference step

i — 1 2. 3

k — 1. 2. 6

The probability model for the conditions listed in table 3-1 is similar to the linear model for a factorial design in experimental design and analysis; i.e.,

$$P(X_j \mid X_1, X_2, \ldots, X_{j-1}) = P_0 + \frac{\Delta P_1 Y_1 + \Delta P_2 Y_2}{\text{Equipment}} + \frac{\Delta P_3 Y_3 + \Delta P_4 Y_4}{\text{Activity}} + \frac{\Delta P_5 Y_5}{\substack{\text{Time} \\ \text{Constraint}}}$$

$$+ \Delta P_6 Y_1 Y_3 + \Delta P_7 Y_1 Y_4 + \Delta P_8 Y_1 Y_5 + \Delta P_9 Y_2 Y_3 + \Delta P_{10} Y_2 Y_4$$

$$+ P_{11} Y_2 Y_5 + P_{12} Y_3 Y_5 + P_{13} Y_4 Y_5 + P_{14} Y_1 Y_3 Y_5 + P_{15} Y_1 Y_4 Y_5$$

$$+ \Delta P_{16} Y_2 Y_3 Y_5 + \Delta P_{18} Y_2 Y_4 Y_5 + \epsilon . \tag{5}$$

Note that equation 5 is linear in terms of the ΔP's and is referred to as a linear model for this reason.

Definitions for terms in equation 5 are as follows:

$P(X_j \mid X_1, X_2, \ldots, X_{j-1})$ = the conditional probability that the reference step is performed correctly, given that the $j-1$ earlier steps have been performed correctly.

Y_1 = 1 if equipment design feature D_1 is present in earlier steps[4]

Y_1 = 0 if equipment design feature D_1 is absent in earlier steps

Y_2 = 1 if equipment design feature D_2 is present in earlier steps

Y_2 = 0 if equipment design feature D_2 is absent in earlier steps

Y_3 = 1 if operator activity A_1 is present in earlier steps

Y_3 = 0 if operator activity A_1 is absent in earlier steps

Y_4 = 1 if operator activity A_2 is present in earlier steps

Y_4 = 0 if operator activity A_2 is absent in earlier steps

Y_5 = 1 if time constraint C_1 is present in earlier steps

Y_5 = 0 if time constraint C_1 is absent in earlier steps

P_0 = mean probability of success in step j when only conditions A_0, D_0, and C_0 are present

ΔP_1 = mean increase or decrease in $P(X_j \mid X_1, X_2 \ldots X_{j-1})$ when D_1 is present in earlier steps

ΔP_2 = mean increase or decrease in $P(X_j \mid X_1, X_2 \ldots X_{j-1})$ when D_2 is present in earlier steps

ΔP_3 = mean increase or decrease in $P(X_j \mid X_1, X_2 \ldots X_{j-1})$ when A_1 is present in earlier steps

---

[4] Values of 1 or 0 assigned to the Y's in the linear model refer to the presence or absence of a variable and not to success or failure in performance of step j.

ΔP_4 = mean increase or decrease in $P(X_j | X_1, X_2 \ldots, X_{j-1})$ when A_2 is present in earlier steps

ΔP_5 = mean increase or decrease in $P(X_j | X_1, X_2 \ldots, X_{j-1})$ when C_1 is present in earlier steps

ΔP_6 = mean increase or decrease in $P(X_j | X_1, X_2 \ldots, X_{j-1})$ due to the interaction between D_1 and A_1

.

.

.

ΔP_{13} = mean increase or decrease in $P(X_j | X_1, X_2 \ldots, X_{j-1})$ due to the interaction between A_2 and C_1

ΔP_{14} = mean increase or decrease in $P(X_j | X_1, X_2 \ldots, X_{j-1})$ due to the interaction between D_1, A_1, and C_1

.

.

.

ΔP_{17} = mean increase or decrease in $P(X_j | X_1, X_2 \ldots, X_{j-1})$ due to the interaction between D_2, A_2, and C_1

ϵ = error in estimating $P(X_j | X_1, X_2 \ldots, X_{j-1})$.

The term P_0 in equation 5 was defined to be the mean probability of success when only the conditions in cell 1 of the matrix are present. By definition, these are the conditions to which the probabilities in the data store apply. Therefore, P_0 may be estimated by means of the appropriate probability value from the data store.

Other parameters in equation 5 represent effects of conditions present in earlier steps in the operational task. Since these conditions are not covered by the data store, one must be concerned with the means for obtaining estimates of their values. Again, one must turn to the methods of experimental design and analysis for an answer. The conditions in table 3-1 are arranged in a factorial design. Equation 5 is a linear model for this design. Therefore, parameters in the model may be estimated from the results of a properly designed experiment. Note, however, that the response or dependent variable in the present instance is a probability value. One estimates probabilities by means of frequency of success values observed in operational situations or in experimental studies. To obtain frequency of success values, one must conduct not one but a series of observations or experiments.

Only a small number of the pertinent variables is included in equation 5. Inclusion of a larger number or of all the pertinent variables obviously would greatly increase the number of terms in the model. In a conventional experiment, a minimum of one observation must be taken for each parameter in the model. If the number of parameters is large, the

work required in conducting a conventional experiment could be excessive. If a series of observations must be taken for each parameter, the work involved will increase accordingly.

Only qualitative variables were included in the example used in developing the linear model (equation 5). When quantitative variables are also included, the model can be constructed to take nonlinear effects into account. For example, if one wished to investigate the effect of several levels of temperature, terms in the model for main effects would take the form,

$$\Delta P_j Y_j + \Delta P_{j+1} Y_j^2 + \Delta P_{j+2} Y_j^3 + \ldots \ldots \tag{6}$$

In expression 6, the variable Y_j takes on the pertinent values of temperature. The parameter ΔP_j represents the average linear effect of temperature; ΔP_{j+1} and ΔP_{j+2} represent the average quadratic and cubic effects, respectively, of temperature.

Discussion

Two major problems require solution before significant progress can occur in the development of transition models: (1) the factors responsible for dependent relationships among steps in a task must be fully identified, and (2) the effects of the dependent relationships (i.e., the ΔP's in the linear model) must be determined. Obviously, the problem of identifying the factors responsible for the dependent relationships must be solved first. Factors cannot be included in the transition model if they are unknown. Equally important is the need to eliminate factors not having a measurable effect, so that the number of terms in the transition can be reduced to manageable proportions.

Success or failure in the development of transition models ultimately may hinge upon the number of interaction effects occurring in the transition model. In the absence of any interaction among factors, it is possible to isolate factors and determine their effects individually. To determine interaction effects, however, one must study factors in combination with one another. As equation 5 demonstrates, a very small number of factors can generate a large number of interactions. However, if one is willing to neglect the higher order interaction terms, the ΔP's may be determined for the main effects and lower order interaction effects by conducting a series of experiments where only a small number of variables are examined in any one experiment.

Summary

Dependent relationships among steps in a task performed by the same operator or by operators working together make it necessary to use conditional probabilities in the model for computing human performance reliability. The value of the conditional probability for a given step depends not only on the characteristics of the equipment being operated and the environment in which the step is performed but also upon the particular combination and characteristics of task steps preceding it in the operational sequence. Sources of probability data available now or likely to become available in the future can take equipment design features and the environment into account. The combination of characteristics of earlier task steps,

however, usually is unique. Consequently, transition models are needed to bridge the gap between the marginal probabilities found in data stores and the conditional probabilities of dependent models for computing human performance reliability. The form of the transition models is similar to that of the linear models of experimental design and analysis. Before significant progress can be made in the development of the models, however, two major problems must be solved: (1) the factors responsible for the dependent relationships among steps in a task must be fully identified, and (2) the effects of the dependent relationships upon probability of successful performance of given task steps must be determined. Success or failure in the development of transition models ultimately may hinge upon the extent to which interaction of factors forming the dependent relationships enter into the transition model.

4. TOWARD A GENERAL CHARACTERIZATION OF ELECTRONIC TROUBLESHOOTING[1]

Anthony K. Mason and Joseph W. Rigney
Electronics Personnel Research Group, University of Southern California

Roughly speaking, corrective maintenance tasks can be classified into those which are accomplished by following a pre-established plan and those which are guided by taking into account information obtained at each step in the troubleshooting process. With regard to this latter category, some recent work[2] was directed toward investigating the resemblance of technician troubleshooting behavior to that of a Bayesian processor.

In the course of the investigation, experiments were performed to compare the decisions reached by human technicians with those implied by the application of Bayes theorem. These decisions were for the isolation of hypotheses concerning the actual circuit malfunction. Analysis of the data obtained from these experiments indicated that although the Bayesian model was reasonably predictive,[2,3] it would be desirable to define a more generalized concept of a troubleshooting processor. In particular, a concept of a processor seemed to be needed which accommodated a number of types of errors that the electronics technicians were making during the troubleshooting procedure.

The purpose of this paper is to present some preliminary suggestions for such a processor, and, in particular, one which accommodates certain categories of error in human electronic troubleshooting.

Relevance to Man-Machine Effectiveness Analysis

There are several factors underlying the desire to formulate a troubleshooting processor model which accommodates a fairly broad spectrum of possible specific procedures. One of these is that such a model would hopefully unify the many alternative ways of characterizing and explaining the troubleshooting behavior of the human technician. Another reason is that if the model does accommodate a broad spectrum of strategies and procedures, it would serve as a vehicle for formulating the cost effectiveness structure associated with the troubleshooting of electronic equipment.

[2] Bond, N. A., Jr., and Rigney, J. W. 'Bayesian Aspects of Trouble Shooting Behavior." Human Factors, Vol. 8, No. 5, October 1966, pp. 377-383.

[3] Rigney, J. W., Cremer, R. H., Towne, D. M., Bond, N. A., Jr., and Mason, A. F. Measurement and prediction of cognitive loadings in corrective maintenance tasks: I. A Bayesian approach. Los Angeles: Univ. Southern California, Electronics Personnel Res. Group, February 1966 (Tech. Rep. 46).

Ideally, it would be of great utility to have a characterization of the electronic troubleshooting process which accommodates the many alternative troubleshooting processors that may be implemented -- whether they be automatic or manual, theoretically optimal or suboptimal, reliable or unreliable. Ideally, such a characterization should be sensitive to the degree of automation that may be introduced in performing troubleshooting tasks. Thus, measures of effectiveness for particular equipments could be generated across the automated to manual spectrum of alternative troubleshooting processors and serve to improve the sensitivity of the "maintainability" component of cost effectiveness models for electronic systems.

The effectiveness of the troubleshooting tasks in system maintainability is influenced by many factors which also influence other aspects of system operation and cost. The efficiency of the human processor as a troubleshooter is influenced by his training which includes knowledge of the specific equipments, fundamental concepts in symptom-malfunction relationships in electronics circuits, and by troubleshooting aids. The troubleshooting aids themselves may be automated. In addition, front panel layout, modularization of the equipment, and a multitude of hardware design considerations have a substantial impact on the efficiency with which an equipment malfunction may be diagnosed. These and many other considerations combine to establish the inherent maintainability of electronic equipment. Until some ultimate diagnostic automata is established, the human processor must be considered to be an alternative within a cost effectiveness analysis.

The intent of the following discussion is not to present explicit cost effectiveness relationships between system characteristics and the diagnostic processor. Rather, it is to consider a characterization of the troubleshooting process which accommodates a number of hypothetical troubleshooting processors. By troubleshooting is meant that portion of the maintenance process which is concerned with isolating the malfunction in the system. That is, it is assumed the troubleshooting process takes as a point of departure the fact that there exists a malfunction and terminates when a decision regarding the malfunction has been made. The next step is to take specific corrective action such as replacing the faulty component.

Electronic Troubleshooting as a Problem Solving Process

It seems reasonable that the process of troubleshooting electronic equipment may be viewed, in a general way, as a problem solving process. For this reason, a very general theory of problem solving should accommodate the specialization of troubleshooting electronic equipment.

Meserovic[4] and others have suggested that ultimately the task of solving a problem may be viewed as the mapping of two sets. This mapping is suggested by the function

$$T(Z:X) \quad Y \qquad\qquad (1)$$

[4] Mesarovic, Mihajlo D., "Toward a formal theory of problem solving," Computer Augmentation of Human Reasoning (Margo A. Sass and William D. Wilkinson, Eds.). Washington, D.C.: Spartan Books, Inc., 1965, pp. 37-64.

where $X = \{x_i\}$ is termed the input set, $Z = \{z_i\}$ is termed the state set,
$Y = \{y_i\}$ is termed the output set, and T is a system transformation.[5]

To relate equation (1) to the problem of diagnosing a system mal-
function, let the set Z characterize the knowledge of the system; let X
denote new information that is obtained by performing some test (taking data)
on the equipment; then the set Y represents the knowledge of the malfunction
following the task. The transformation T is the way in which previous
knowledge of the equipment and new information is processed or modified to
obtain Y.

For the process suggested in equation (1) to be of utility, it is neces-
sary to explicitly define the input set, the state set, and the properties of
the transformation of these sets in terms of electronic troubleshooting.

A Hypothesis Space for Electronics Troubleshooting

The "state" of the troubleshooting problem may be characterized by
a hypothesis space. The points in this space are the possible malfunctions.
We may denote this space by a set S,

$$S = \{h_1,\ h_2,\ \ldots,\ h_n\}$$

where h_i is a hypothesis regarding what is wrong with the equipment. For
troubleshooting at the circuit level in terms of single, catastrophic failures,
it is convenient to think of h_i as the hypothesis that the ith component in the
circuit is the malfunction. However, the elements of S may be viewed as
hypotheses regarding the reason for system malfunction at any level of
system troubleshooting and regardless of whether dealing with system
degradation or catastrophic failure, single or multiple component failures.
For purposes of providing examples, the following discussion focuses on
troubleshooting situations in which the hypothesis space denotes single
catastrophic failures among n components in a circuit.

For those troubleshooting procedures which are guided by taking into
account information obtained at each step in the troubleshooting process,
the troubleshooting processor makes a sequence of tests on the electronic
equipment. These diagnostic tests, for example, detecting an abnormally
high voltage at a certain test point, are used by the processor to partition
the hypothesis space into subsets which contain relatively true and
relatively false hypotheses. Thus, a particular diagnostic test may be
used to modify the problem by modifying the hypothesis space. The process
is repeated until the processor specifies the malfunction or is unable, on
the basis of its understanding of the electronic system, and available tests
to reach a decision.

The processor may make errors in several categories. These include
incorrectly taking the test reading, incorrectly interpreting the test reading,
and incorrectly modifying the hypothesis space. More specific types of
errors may be defined within each of these categories. As a result of these
errors, the wrong hypothesis may be selected. A processor which is
correctly taking data, correctly interpreting data, and correctly adjusting
the hypothesis space will correctly isolate the circuit malfunction if a

---

[5] The transformation, T, may simply cause a reformulation of the original
problem.

sufficient base of symptom-malfunction information is available. The human technician not only makes errors in all of these categories, but may have an insufficient information base. This procedure is made more explicit as follows:

Denote by S_i those points in the hypothesis space, S, which, as a result of the ith test, are still possible hypotheses regarding the circuit malfunction. S_i is obtained from S by a transformation which is denoted

$$T(S: D_i, \ t_i) = S_i \tag{2}$$

and where

S = the original problem hypothesis space

S_i = those hypotheses in S which are true as a result of the ith diagnostic test (note that the subscript i denotes the sequence number of the test rather than a test identification number).

t_i = the electronic reading obtained at the ith test made (for example: t_i might be 100 volts, 0 ohms, "a high voltage," etc.).

D_i = an ordered set of reading associated with each hypothesis for the ith test made: $D_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,n}\}$ where each element $d_{i,j}$ is the reading at the ith test given hypothesis j is the malfunction.

Since S_i is the set of hypotheses with test readings at the ith test made that corresponds to the elements of D_i, it follows that S_i may be defined as

$$S_i = \{\text{all } h_j \epsilon S \ \overset{4}{\text{such that}} \text{ for each } j, \ D_i \ni d_{i,j} = t_i\}.$$

Thus, the transformation T is one of matching the test reading t_i against the symptom-malfunction relationships expressed in D_i to partition a set, S_i.

Denote by $S_k = \{S_1, \ S_2, \ \ldots, \ S_k\}$ the family of sets which are the possible malfunctions as a result of each of the individual k tests made. There are operations on S which characterize the behavior of the processor in attempting to isolate the malfunction. For instance, denote by M_k the intersection of the elements of the members of S_k. That is

$$M_k = \underset{k}{\cap} \ S_k = S_1 \cap S_2 \ \ldots \cap S_k.$$

Now M_k may be viewed as the set of true hypotheses in S as a result of a sequence of k tests. Note that S_i is the set of true hypotheses on the basis of the ith test only. Consider the following example:

There are 9 possible malfunctions in the circuit. Let $S = \{h_1, h_2, h_3, \ldots, h_9\}$; the hypothesis space, S, is illustrated in Figure 4-1A. Suppose that the first test made yielded a result of 0^V, that is $t_1 = 0^V$, and that $D_1 = \{0^V, 100^V, 300^V, 0^V, 0^V, 100^V, 300^V, 100^V, 0^V\}$. The elements of D_1 are the expected test readings given each of the 9 malfunctions. Note that $d_{1,1} = 0$ volts means that given malfunction 1, the expected test reading for the 1st test made is 0 volts. Then $T(S: D_1, t_1) = S_1 = \{h_1, h_4, h_5, h_9\}$ as suggested in Figure 4-1B.
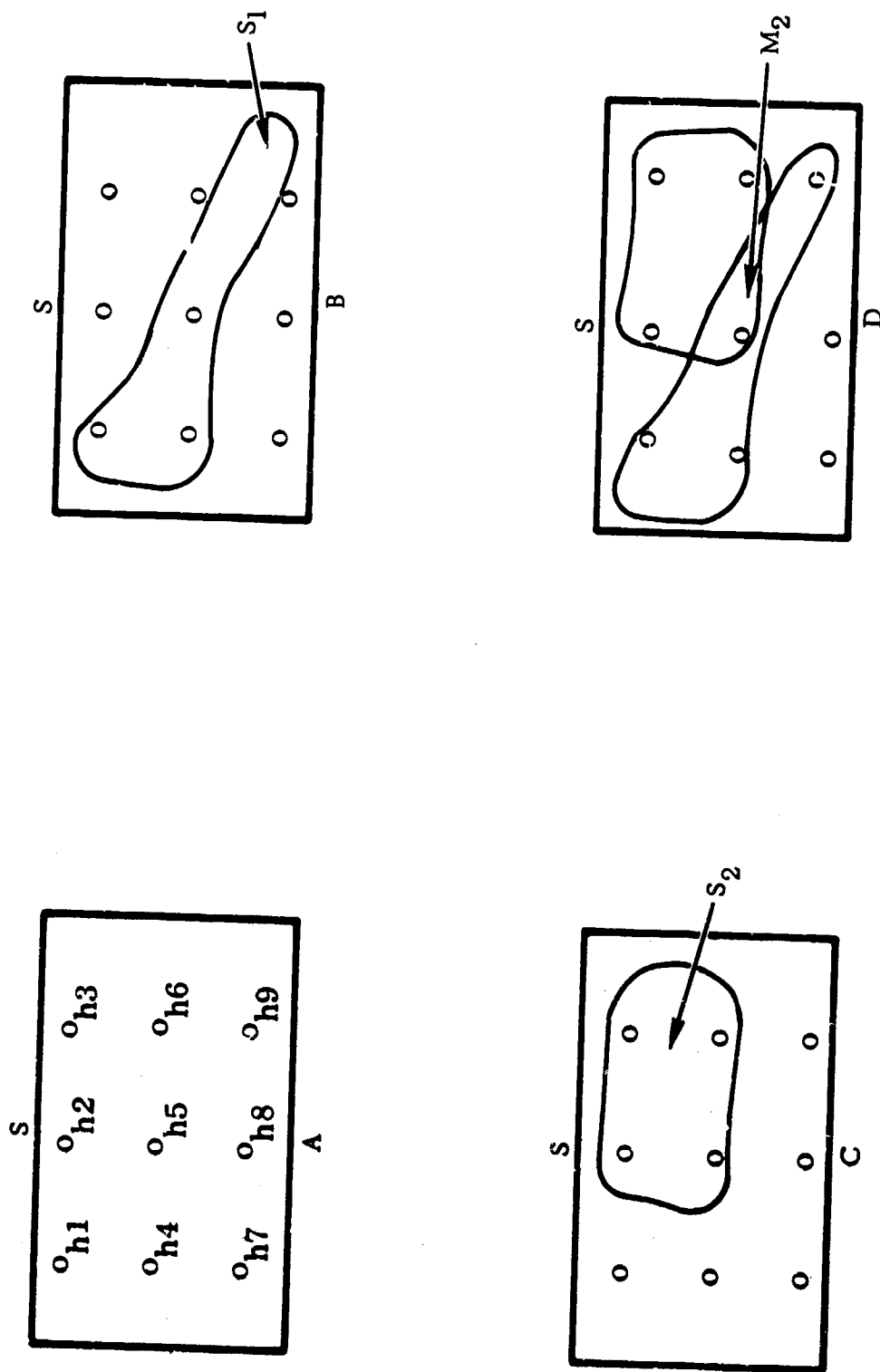
Figure 4-1. The Hypothesis Space for the Example

Suppose the result of the second test is "low." Then t_2 = "low.' In addition, assume D_2 {high, low, low, high, low, low, normal, high, normal}. An element of D_2 which is "low" means that given the malfunction, the reading at that test point will be low relative to its normal value.

Then $T(S; D_2, t_2) = S_2 = \{h_2, h_3, h_5, h_6\}$ as illustrated in Figure 4-1C. Now $M_1 = S_1$, but $M_2 = S_1 \cap S_2$, and is diagrammed as shown in Figure 4-1D. It should be kept in mind that M_k is the result of just one of many operations that may be defined on the family S.

Without varying the transformation T, equation (2) may be extended by defining some new arguments for the function. For example:

$$T(M_{k-1}; D_k, t_k) = M_k = \bigcap_k S_k = S_1 \cap S_2 \ . \ . \ \cap S_k \qquad (3)$$

Equation (3) indicates that if M_{k-1} is substituted for S in equation (2), we have characterized a processor which is a perfect processor in the sense that it is using all previous test results to reduce the hypothesis space. On the other hand, a processor which is always operating with arguments

$$T(S; D_i t_i) = S_i$$

represents a processor which is using no previous results for the reduction of the hypothesis space. To characterize a processor which is using the results of some but not all previous tests, let

$$S_{(m, k)} = S_{k-1} \cap S_{k-2} \cap \ldots \cap S_{k-m}$$

That is, $S_{(m, k)}$ is the set of hypotheses which remain as the result of the m previous tests. Then

$$T(S_{(m, k)}; D_k, t_k) = S_{k-m} \quad S_{k-m+1} \cdots S_k. \qquad (4)$$

The motivation for equation (4) includes the fact that some preliminary experiments rather clearly suggest that the human technician is operating on a hypothesis space which is reduced according to the results of the last couple of tests. It also may be noted that letting $S_{(0, k)}$ = S, equation (4) reduces to equation (2) and by letting m k-1, equation (4) reduces to the perfect processor suggested in equation (3).

Although a function T with various arguments may be used to provide a specification of the way in which the processor modifies the hypothesis space, it does not specify when a diagnostic decision will be made or what tests will be used in the test sequence. There are measures on the hypothesis space that may be used to answer these questions.

Assume for the moment that the diagnostic test data is taken without error: that is, t_i is accurate. Further suppose that the set D_i, the symptom-malfunction relationships for test i, are accurate and deterministic in the sense that $P(d_{i, j} \ t_i h_j)$ 0 or 1.[6] That is, the test data either does match or does not match the known symptom data.

_____

[6] $P(d_{i, j} \ t_i h_j)$ is read "the probability that $d_{i, j}$ t_i given h_j'; it is the probability of the test data given the hypothesis.

The hypothesis space may be mapped into a probability space using Bayes theorem.

$$P(h_j, t_i, D_i) = \frac{P(h_j) \, P(d_{i,j} \mid t_i \, h_j)}{\sum_j P(h_j) \, P(d_{i,j} \mid t_i \, h_j)} \tag{5}$$

Before any diagnostic test is taken, a priori probabilities may be assigned to the n hypotheses according to their a priori probability of being the malfunction. Thus, if the probabilities that the test data will match the known symptom hypothesis relationships are all 0 and 1, the repeated application of equation (5) will eventually reduce the probabilities of the hypotheses to zero except for a single hypothesis with probability 1. For this to occur, it is necessary that sufficient sets D be available and that they be consistent. If $P(d_{i,j} \mid t_i \, h_j)$ is not equal to 1 or 0, the hypothesis space is partitioned into sets which represent hypotheses that are more or less likely to the true hypothesis.

In addition, an information content measure may be defined on the hypothesis space. The "information level" of the troubleshooting task at the ith diagnostic test is defined by the well known function

$$H_i = -\sum p(h_j) \, \log_2(p(h_j)). \tag{6}$$

The next diagnostic test may then be specified as the test which causes the greatest reduction in H, the information level. In other words, the ith test should be the test which maximizes the expression $(H_{i-1} - H_i)$. Since $p(h_j)$ is calculated using equation (5) and t_i is unknown before making a test, the decision rule may be stated as

$$\underset{D}{\text{MAX}} \left[H_{i-1} - H_i \right].$$

This rule may be used to generate a sequence of diagnostic tests which minimizes the number of tests required to isolate a malfunction. In using this procedure, the probability space resulting from the application of equation (5) must be implicitly determined for each possible next test so that an optimum test can be selected. This procedure is a generalization of the well known "half-splitting" strategy for the isolation of circuit malfunctions.

Now these relationships serve to define a very efficient troubleshooting processor. In particular, equation (3),

$$T(M_{k-1}, D_k, t_k) = M_k$$

which defines the way in which a processor utilizes all previous tests and the current test data to modify the hypotheses space; the use of Bayes theorem, equation (5), for the development of a probability space at each diagnostic step which can be used to elicit a decision as to which hypothesis is correct; and the use of an information measure, equation (6), to dictate a next most efficient step. As a practical matter, however, the realization of such a processor is hindered by a number of considerations.

(1) With regard to the making of a test reading t_i, it is assumed that a human technician must make a set-up on the equipment, properly connect test equipment, and take a visual or audio reading. In automatic test equipment, a sensor imbedded in some stage of the equipment is required to take the reading. In either case, t_i may be in error. There are several possible consequences of an error in t_i. These include rejecting correct hypotheses of malfunctions and/or accepting incorrect hypotheses of malfunctions. Each of these possiblities may be characterized on the hypothesis space. With an electronic technician, this type of error either leads to an incorrect decision as to the actual circuit malfunction or results in confusion over the state of the equipment which sometimes leads to giving up the task. Clearly, the consequences of making a mistake in obtaining t_i depend not only on the symptom-malfunction relationships contained in D, but on the state argument being used in the processor T.

(2) With regard to the symptom-malfunction relationships suggested in the set D, it is assumed that the human technician has acquired these as a result of training in fundamental symptom-malfunction relationships; has acquired a "feel" for them as a result of troubleshooting experience on the equipment; or has them provided to him in the form of troubleshooting software aids. With regard to automatic test equipment, the symptom-malfunction relationships are normally found in computer memory. As above, errors in D_i cause the processor to accept or reject hypotheses incorrectly and the state of the processor can be depicted on the hypothesis space of the task. Normally, $P(d_{i,j} = t_i | h_j)$ is _not_ equal to 0 or 1; that is, the symptom-malfunction relationship is not deterministic but is probabilistic. This is because explicit system operating characteristics are difficult to define. In addition, if the processor is a human technician, he is simply not sure of the symptom-malfunction characteristics of the circuit or system.

(3) With regard to the hypothesis space that is used for an argument in T, automata are capable of accurately identifying and carrying large spaces of this nature. On the other hand, experiments would indicate that the human technician works with not more than 4 or 5 points in this space at any one time while troubleshooting at the circuit level. Thus, the hypothesis space is partitioned by the technician at the outset, and the search for a true hypothesis is exhausted before another subset is focused upon.

(4) The use of Bayes theorem as indicated in equation (5) as a model of the decision made by the human technician has been experimentally checked in terms of the total hypothesis space, S. That is, equation (5) was applied under a processor of the form $T(M_{k-1}; D_k, t_k)$. The arguments D_i were obtained by determining the technicians' understanding of symptom-hypothesis characteristics of simple circuits. In addition, the efficiency of the test sequence elected by the technician was measured in terms of the optimum test sequence of information content reduction. It is believed that modification of the state argument under T will substantially improve the predictability of these models.

Some Planned Experiments

In order to better characterize the hypothesis space which is used by the human troubleshooter, several preliminary experiments are planned. These experiments involve the use of the test console shown in Figures 4-2 and 4-3. There are two display panels shown. One presents information to the subject, S, and one displays the current hypothesis space of the subject and displays the sequence of actions taken by the subject.
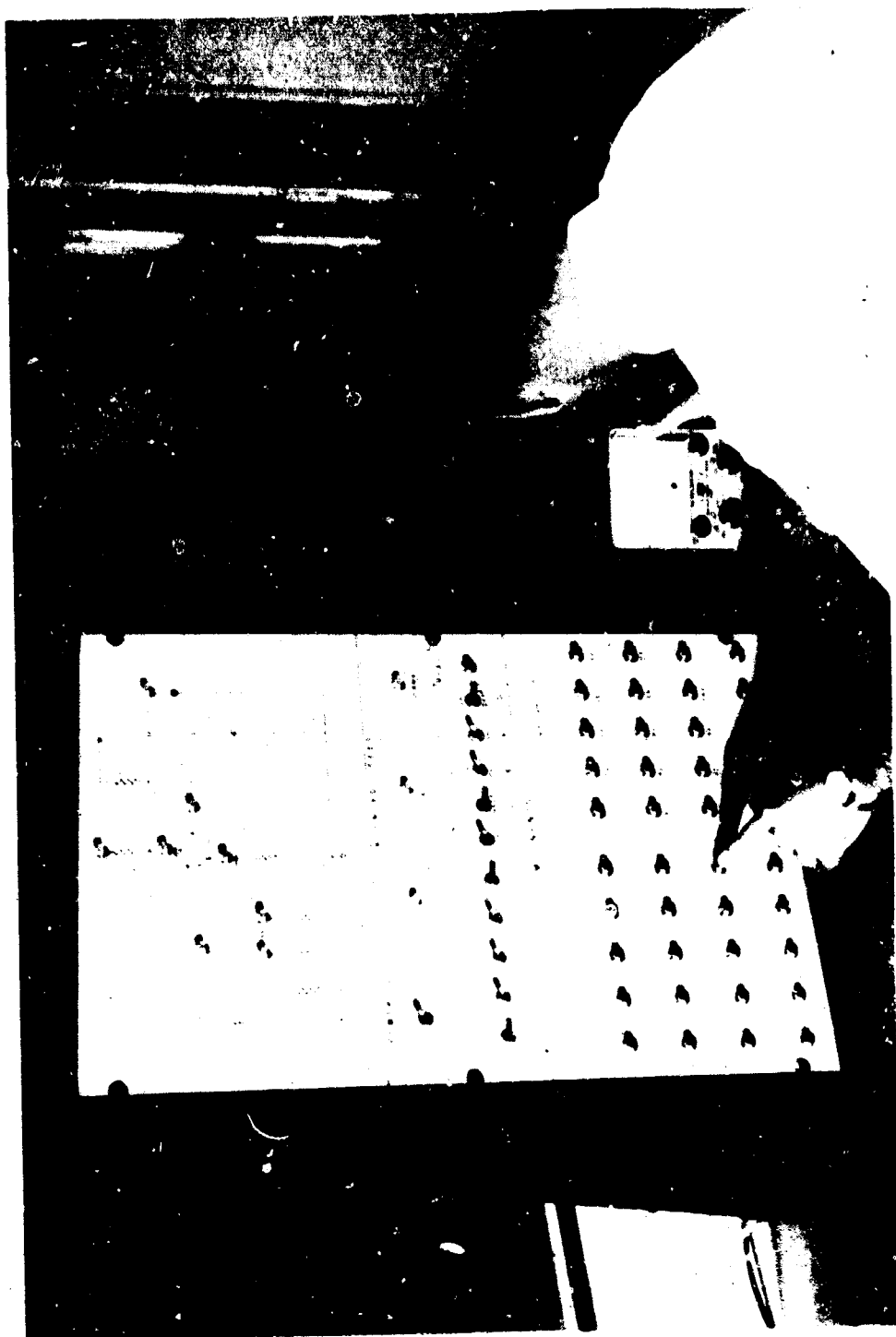
Figure 4-2. The Subject Test Panel

Figure 4-3. The Test Apparatus Including the Subject Panel, Panels Displaying Subject Hypothesis Space, and Video Recording Equipment

At the top of the subject panel is a schematic of an electronic circuit. Test points are identified at various points on the circuit schematic. These test points are actually button switches which are actuated by the subject to take a reading. S may take DC, AC and Ohm readings at each of the test points. A multimeter is connected to a terminal on the front panel. To properly take a test reading, S must set-up the toggle switches denoting which reading is intended, put the meter in the proper mode, and depress a button on the circuit schematic. S verbally describes to the test monitor what the expected reading should be before the test is taken and what reading was observed.

Each of the possible malfunctions in the circuit is associated with a pair of buttons on the bottom of the panel. At the conclusion of each test, S depresses buttons to indicate which hypotheses he feels are no longer under consideration and which hypotheses he feels still may be possible. All information is displayed on the monitor panel and is recorded on video tape. This allows a permanent record of the sequence of actions, errors, and time at which actions were performed. It facilitates the calculation of the interval of time between certain tasks (the frame counter on the video tape recorder is used to record cumulative time).

The experiments performed to date with this equipment have been generally along the following lines. S is told that there is a malfunction in the circuit. S proceeds to make a diagnostic test by taking a reading. While making this reading, he has an opportunity to make errors in setting up the front panel, setting up his test equipment, and in observing the test reading. He then is urged to make some assertion concerning the nature of his hypothesis space by pressing buttons to indicate which hypotheses he thinks may be true and which false. No change in the hypothesis space may be a response. Once S makes a diagnostic decision, say, replace a resistor, E switches in a good component to effect the replacement. S then proceeds to verify that his diagnostic decision was correct. Before concluding, S is required to assert that the circuit is now in normal operating condition.

All symptom-hypothesis relationships for the circuit are known. In addition, it is possible to have S take a paper-and-pencil test which allows the construction of his initial concept of symptom-malfunction relationships in the circuit. That is $P(d_{i, j} = t_i | h_j)$ are obtained according to the technician's understanding of the circuit. Some experiments in this area have indicated that there are substantial changes in S's symptom-malfunction concepts during the actual diagnostic process: generally, S benefits from the exercise, and an improvement in the quality of the symptom-malfunction relationships in the circuit may be detected. The effect of this, of course, is that the sets D are not constant throughout the experiment. It is hoped that further information on changes in D during the diagnostic procedure will be obtained as a result of S's indicating the expected reading at each test as he proceeds.

The central motivation for the experiments lies in obtaining an improved understanding of the hypothesis space that is used by the technician, and, ultimately, an improved model of the human technician as a processor.

5. THE SANDIA HUMAN ERROR RATE BANK (SHERB)

Lynn V. Rigby
Sandia Corporation

The Sandia Human Error Rate Bank (SHERB) is not exactly an accomplished fact. It is something we have planned for a long time, and do work at occasionally, but it is still merely a small number of file cards contained in a small file box, plus a few rough notes and data not yet transferred to the cards. Nonetheless, we felt that the philosophy, methodology, and experience behind the file and the format used for the file would be of value to anyone with similar interests.

Background

Such a data bank is by no means an original idea. You are doubtlessly aware of the Index of Electronic Operability Data Store developed by the American Institutes for Research[1]. This is still the most comprehensive listing of human errors available, but the literature contains many other compilations of human error rates, such as the very useful lists compiled by Dunlap and Associates[2], Aerojet-General[3], General Electric[4], and Rocketdyne[5].

Other listings and pertinent data can be found in a wide variety of sources, such as industrial engineering works, quality control reports, safety reports, and the general psychological literature. In fact, SHERB actually began some years ago as a contract with the University of New Mexico in which, in essence, Sandia asked psychology graduate students to search the literature for records of human error rates in production tasks[6]. That preliminary study led to a larger effort, again with the support of the University of New Mexico, and we soon hope to publish a 5000-item bibliography of sources of human performance, and particularly human error. data. This bibliography is now being indexed.

[1]Munger, S.J., Smith, R.W. and Payne, D., An index of electronic equipment operability: data store. Pittsburgh, Pa.: American Institutes for Research Report AIR-C43-1/62-RP(1), January 1962.

[2]Mitchell, M.B., Smith, R.L. and Verdi, A.P., Development of a technique for establishing personnel performance standards (TEppS): Phase III - final report. Santa Monica, Calif.: Dunlap & Assoc., Inc., July 1966.

[3]Irwin, I.A., Levits, J.J. and Freed, A.M., Human reliability in the performance of maintenance. Sacramento, Calif.: Aerojet-General Corp. Report LRP 317/TDR-63-218, May 1964.

[4]Stave, A.M., The quantification of human reliability, a feasibility demonstration. Philadelphia, Pa.: General Electric Spacecraft Department Report TIS 65SD216, March 1965.

[5]Peters, G.A., Hall, F.S. and Kuplent, C., Human reliability data. Canoga Park, Calif.: Rocketdyne Report IDEP 347.90.00.00-G1-03, June 1965.

[6]Hurlock, R.E. and Peterson, G.M., A survey of the literature on human error. Albuquerque, N.M.: University of New Mexico, January 1963.

Concurrent with the bibliographic effort. we collected copies in various forms of some 3000 reprints of itsms listed in the bibliography. These reprints are now on microfilm indexed for quick access. The ultimate goal was. and still is. to convert the usable data in all those documents into a common and easily accessible data file. now called SHERB. Due to the pressures of higher priority tasks. this effort is proceeding slowly. but it is proceeding.

Why SHERB?

Before discussing the file itself. it may be well to consider the basic question. Why SHERB? The human factors group at Sandia is part of the Systems Reliability Division, and its primary purpose is to quantify human performance contributions to system reliability. In order to be meaningful. such quantification must be compatible with common reliability statistics, and the one aspect of human performance that is compatible is human error.

If human error is defined to be any variant of human performance that reduces the probability of system or mission success. then failures due to human error can be treated in a manner very similar to component failures; that is. human errors can be predicted as a probabilistic function of the variables determining or influencing that human performance related to system performance.

The prediction techniques employed at Sandia have been described by Rook[7, 8, 9] and Swain[10, 11, 12, 13]. These techniques depend primarily upon a detailed functions and task analysis; the preparation of logic tree diagrams to allow analysis of the relevant inputs. outputs. interactions. pertinent variables. and consequences; the estimation of the probability associated with each limb of the tree diagram; and the appropriate probability statistics.

---

[7]Rook. L.W., Reduction of human error in industrial production. Albuquerque. N.M. : Sandia Corporation Technical Memorandum SCTM 93-62(14), June 1962.

[8]Rook. L.W.. Evaluation of system performance from rank-order data. Human Factors. 1964. 6. 533-536.

[9]Rook. L.W.. Motivation and human error. Albuquerque. N.M. : Sandia Corporation Technical Memorandum SC-TM-65-135. September 1965.

[10]Swain. A.D.. A method for performing a human factors reliability analysis. Albuquerque. N.M. : Sandia Corporation Monograph SCR-685. August 1963.

[11]Swain. A.D.. THERP. Albuquerque. N.M. : Sandia Corporation Reprint SC-R-1338. August 1964.

[12]Swain. A.D.. Field calibrated simulation. Albuquerque. N.M. : Sandia Corporation Reprint SC-R-67-1045. February 1967.

[13]Swain. A.D.. Some limitations in using the simple multiplicative model in behavior quantification. Reliability of human performance in work: a symposium of the 1966 annual convention of the American Phychological Association. Wright-Patterson Air Force Base. Ohio: Aerospace Medical Research Laboratories Technical Report AMRL-TR-67-88. in press.

In any human task, a large number of discrete inputs, outputs, and influencing variables come into play; and the human error analyst must be able to assign occurrence and error probabilities to <u>all</u> of those that can effect system failure. Despite our preferences for scientific rigor, there is seldom time or funds to conduct experiments to obtain situation-specific data; so we must depend, and depend heavily, upon our ability to extrapolate from the known to the unknown, however unlikely the two may be.

SHERB, past experience, and whatever can be found in a quick look at the literature constitute our pool of knows for any given application. It is an inexact and heterogeneous pool and, despite care and expertise in interpretation, our predictions can be considerably in error. But though accuracy is to be desired and sought, inaccuracy is no bar to our efforts.

Whenever we feel strongly enough about an error-likely situation to make an issue of it, we find others easy to convince that human error is so important that gross predictions are better than none. Usually, no one is really concerned with the accuracy of our figures, yet almost everyone is willing to listen <u>if we have figures</u>; and they are willing to accept the figures as reasonable once the basis and implications are presented. Such experience merely underscores three common expectations:

1. Scientists and engineers fully expect human performance to have a large impact on system performance; they need only to be shown <u>how</u> and to what <u>degree</u>.

2. Numbers <u>are</u> the fundamental structure of any decision fabric in any scientific and engineering environment.

3. The contribution of a human error analyst is primarily dependent upon <u>how quickly</u> he can produce relevant and acceptable estimates.

Thus, the more data we have in SHERB, the larger our pool of "knowns," the better qualified we are to make predictions, the more confidence we have in those predictions, the more work situations we can address, and the more frequently and more quickly we can contribute to a fuller and more accurate interpretation of system success or failure.

The SHERB Format

As it now stands, SHERB consists of a number of 5 x 8 inch file cards. These cards are pre-printed in the format provided in Figures 5-1 and 5-2, which show the front and back sides, respectively. Data are entered upon the cards by hand or typewriter, and the cards are filed alphabetically by task. The number of cards is small, but will increase in time; and as the file grows, more sophisticated filing and cross reference systems can be readily applied, but these are not yet necessary.

In using the file, we simply flip through the cards until we find data appropriate to the task or error we are interested in. If there is more than one card for that task or error, we must decide which set of data is most appropriate (or least inappropriate). If there is no suitable information in the file, we must develop estimates from some other basis. This usually requires some literature search, a paper analysis, and a lot of soul searching. The information on the card ordinarily fills our immediate needs, but the reference can be readily checked for further details and background.

As shown in Figure 5-1, the top of the SHERB card provides for topic descriptions of the interest area, task, type of error, and criterion for error. These blanks are filled with such representative topics as:

Area	Task	Error	Criterion
Assembly	Access	Abuse	Accident
Communication	Checkout	Interchanging	Accuracy
Design	Connection	Mismating	Completion
Inspection	Disconnection	Misreading	Consumption
Installation	Display, linear	Misuse	Cost
Maintenance	Fastening	Omission	Injury
Measurement	Fault diagnosis	Reversal	Man time
Operation	Handling	Substitution	System time

Along the left side of the card shown in Figure 5-1, the basic data descriptors are recorded: these include the mean human error rate, the standard deviation or comparable distribution parameter, the range, and the shape of the distribution, where these can be determined. By human error rate we mean the probability of error per opportunity for error. Such information, of course, allows some latitude in extrapolation. For instance, if the data are applicable to a situation in which other parameters seem notably higher or lower, we may choose some ordinate other than the mean as the basis for prediction. Any such choice is both the exercise and the proof of expertise, but the logic becomes tenuous to the degree that distribution parameters are unknown.

In recording the data, we use whatever significant digits are provided by the source, and leave any rounding to the instance of use, although one significant digit usually reflects the accuracy of the data. The figures are listed as decimals, for example, as 0.0021, rather than 21×10^{-4} or to some standard base such as 10^{-6}. Decimals are more easily grasped and more commonly understood, at least up to five or six decimal places.

SHERB CARD, Sandia Corporation

TASK: Connectors, AN/TRi-Lock* ERROR: QEST Found Defective

AREA: All, Criterion Data CRITERION: QEST

DATA BREAKDOWN:

QEST Deficiencies noted in AN & TRI-Lock Connectors	Number Occurrences	% of Errors	HER
Number of connectors inspected	12,587	--	---
Connectors w/bent pins	19	37%	0.0015
Connectors w/external damage	11	22	0.00087
Connectors improperly mated	9	18	0.0007
Connectors w/parts omitted**	12	23	0.00095
Total connector errors:	51	100%	0.004***

*Based on old type tri-lock, pre scoop-proof design.
**Probably assembly errors.
***p defective connection due to one or more human errors.

REVIEWER: L. V. Rigby ORG. 2152 DATE: 1 Jun 1967

Mean HER:	0.0040
Std. Dev.:	n/a
Lo Range:	n/a
Hi Range:	n/a
Distr. Shape:	n/a
N Opport.:	12,587
Job Area:	Criterion
Kind Data:	Criterion
N Subjects:	n/a
Kind Subjs:	n/a
Work Envir.:	n/a
Climate:	n/a

	-3	-3	0	+1	+2	+3
Task Stress		Varied				
Workspace HE		Varied				
Equipment HE		Varied				
Qual Perf Aid		Varied				
Qual Support		Varied				
Reliability						x
Validity						
Generality			x			
Source Cred.						x

Figure 5-1. The Front Side of a Typical SHERB Card

In the "No Opportunity" blank, we fill in whatever denominator information is provided. This seems to be an inadequately understood area. In any assembly task, for instance, it is not sufficient merely to record the number of soldering errors per number cf units produced. In order to be fully meaningful, the data must show the number of soldering points per unit. at least. It is also helpful to show any differences among the soldering points that might make a difference in either frequency or type of error. For instance, were all wires inserted through holes and soldered, or were some looped, wrapped, or pigtailed?

Similarly, brief topic descriptors are used to identify the job area. the kind of data, the kind and level of subjects. the working environment. and the climatic conditions the data were obtained under. The number of subjects is taken as given in the source, and representative topics in each of the other areas include:

Job Area	Kind Data	Subjects	Work Envir.	Climate
Auto driver	Accident/Incident	Analysts	Airborne	Arctic
Clerk	Deficiency report	Naive	Factory	Desert
Navigator	Feedback data	Task skilled	Field unit	High altitude
Pilot	Field test data	Tech reps	Laboratory	Indoor, Std.
Secretary	Lab experiment	Semi-skilled	Office	Under sea
Technician	A/A inspection	Students	Space borne	Z.I.

Such topics merely indicate the general conditions under which the data were obtained. and the next few rows identify and evaluate the major assumptions underlying the data. particularly:

The stress level the subjects were working under

The quality of workspace human engineering

The quality of equipment human engineering

The quality and representativeness of performance aids used

The quality of supply and support employed or assumed

The above are rated on a seven-point scale via checks made directly on the SHERB card, as shown in Figure 5-1. The values in the scale indicate the following ranges:

-3 = worse than -3σ ($\sim$ worst 0.1%)

-2 = between -2σ and -3σ ($\sim$ 2%)

-1 = between -1σ and -2σ ($\sim$ 14%)

0 = 1σ ($\sim$ 68%)

$+1$ = between $+1\sigma$ and $+2\sigma$ ($\sim$ 14%)

$+2$ = between $+2\sigma$ and $+3\sigma$ ($\sim$ 2%)

$+3$ = better than $+3\sigma$ ($\sim$ best 0.1%)

The use of this kind of scale is not intended to imply greater accuracy in rating; rather, it simply forces us to think in terms of a normal distribution of events. The great majority of events are "more or less average." and they receive the middle, or zero, rating. This kind of rating scale seems to be more useful and more appropriate to probability analysis than a linear scale.

Similar evaluations are made of the statistical reliability (repeatability), validity re the test or experimental situation, generalizability of the data beyond the test or experimental situation, and credibility of the source. Such notes, which are largely subjective, are merely reminders of the general limitations of the data. We may ignore these limitations, but at least we know what they were or seemed to be.

The rest of the card is essentially unstructured. The front allows condensation of any detailed breakdown of the data, as illustrated in Figure 5-1. and the reviewer is identified by name, organization, and date at the bottom of the card. Where others in the human factors group are familiar with the source work, we have them review and corroborate the evaluation.

The back of the card, as illustrated in Figure 5-2. is filled with abstracted narrative in accordance with the following instructions:

1. Task description. What task was being performed when the error was made? How frequently was this task performed? What kinds of activities intervened? What were the task inputs and outputs? And how was the task performed?

2. Error description. What was the nature of the error class or classes? What tolerance limits or requirements defined the error? And what criteria were used in the tabulation of error?

3. Situational variables. In general, what was the situation in which the task was performed and errors made? Were any key independent parameters important to definition or interpretation of errors? Were there conditions which may have systematically increased or decreased the chargeability, detectability, or recordability of errors? Were there any artifactual restrictions which may influence the generalizability of the findings? If there was any analysis or test of significance, show the procedures employed, results obtained, and conclusions drawn.

DESCRIBE: THESE DATA CONSTITUTE ALL CONNECTOR DEFICIENCIES BY QUEST (QUALITY EVALUATION SYSTEM TEST) BETWEEN JAN. 1960 AND AUG. 1961, FOR VARYING NUMBERS OF DIFFERENT KINDS OF NUCLEAR WEAPONS.

DESCRIBE ERROR: ERRORS RECORDED ARE ALL DEFECTS WHICH WOULD LIMIT THE OPERABILITY OF THE CONNECTION. EXCEPT WHERE SHOWN, THESE ERRORS ARE MOST LIKELY ATTRIBUTABLE TO THE LAST INSTALLATION ACTION

DESCRIBE SITUATION: THE DATA LISTED ARE CRITERION DATA IN THAT QEST EXHAUSTIVELY AND SYSTEMATICALLY REVEALS ALL DEFICIENCIES IN THE EQUIPMENT INSPECTED. THESE, THEN, WERE THE ACTUAL AND TOTAL NUMBER OF CONNECTOR PROBLEMS DISCLOSED IN THAT TIME PERIOD. CLASSIFIED DETAILS ARE PROVIDED IN THE SOURCE DOCUMENT.

KEY VARIABLES
RESTRICTIONS:

SOURCE:

Figure 5-2. The Back Side of a Typical SHERB Card

4. <u>Source</u>. Provide a complete bibliographic reference – authors, title, document number, publisher, city and state, date, DDC or other reference number, classification, and page reference.

All of the foregoing matters are completely dependent upon the information provided by the source. If the source does not make such matters clear, we can either estimate the apparent conditions or leave the card blank in that area. In either case, we have just that much less of an idea of how relevant the data are to any potential application. Of course, these are the kinds of information which are, or should be, provided by even reasonably thorough research reporting.

Data Sources and Interest Areas

The data incorporated into SHERB comes from many sources. Most of it is extracted directly from the literature, particularly works already mentioned. Some of it is derived from Sandia development and field tests, some from special Sandia studies (unpublished), and some of it consists of estimates that we have had to develop at one time or another and keep on file for later use. A summary of the major kinds of data encountered, and estimates of their relative merits, is provided in Table 5-1.

TABLE 5-1. Evaluation of Human Error Data Available

Kind of Data	Availability	HER Coverage	HER Reliability	HER Validity
Q/A In-Plant Inspections*	Good	Poor	Poor	Poor
Individual opinion, no analysis	Good	Good	Poor	Poor
Acceptance test data*	Fair	Poor	Fair	Fair
Individual analytic estimate	Poor	Good	Fair	Fair
Accident/Incident data summary*	Good	Poor	Fair	Fair
In Work Deficiency Reports*	Poor	Poor	Fair	Good
Field Feedback Data*	Fair	Poor	Fair	Good
Accident/Incident data raw*	Poor	Poor	Good	Good
Field Test Data*	Fair	Poor	Fair	Good
Mean of Scaled Opinion	Poor	Good	Good	Good
Experiment in Work Situation	Poor	Good	Good	Good
Quality Evaluation System Test	Good	Fair	Good	Good
Laboratory Experiment	Good	Good	Good	Good

*Assuming good denominator information, which is usually lacking.

With the present paucity of such data, we really don't do much in the way of selection. If we can find it, we will use it, at least until better is available. But the information must be convertable to the probability of error per opportunity for error; data which do not have good denominator information are essentially useless, except to indicate failure events or modes. We are, of course, primarily concerned with four broad species of human error:

1. Assembly errors are human errors committed in component and equipment production, which somehow pass acceptance procedures and remain undetected until they cause problems in the field. These include both things like soldering errors, which eventually cause failures outright, and defects which may contribute to other errors, such as an off-center handle or control, etc. Incidentally, we are beginning to believe that undetected assembly error is the primary source of unreliability, particularly in equipment composed of highly reliable components.

2. Installation errors are human errors committed in the installation or integration of a unit into a larger equipment or facility complex. Like assembly errors, installation errors may have long lasting effects on total system reliability, particularly if we include the integration of operational procedures.

3. Operator errors are human errors committed in the operation of the equipment and associated transport, handling or support equipment. The effects of such errors are directly related to both equipment and reliability and mission success or failure.

4. Maintenance errors are human errors committed in the performance of equipment maintenance, which directly influence equipment reliability and thereby indirectly influence mission success or failure. Maintenance can also directly influence mission success.

Taken in aggregate, the above account for a large portion of total system failure. Just how much is a matter of growing concern, and this concern, we hope, will be accompanied by increasing attention to systematic prediction and measurement of human error. Our own experience indicates that the percentage of system failures caused by human error is at least as high as the 50 to 60 percent suggested by the classic studies of Shapero[14] and Zeller[15] and can be as high as 80 to 90 percent in some cases.

Unfortunately, accidents and mission failures resulting from human errors that do not result in equipment failures are not reported with the same regularity and accuracy as equipment failures. And even the reporting of equipment failures omits much good human error data. Our greatest need is still for good feedback data to tell us not only what the real problems are,

[14]Shapero, A., Cooper, J.I., Rappaport, J., Schaeffer, R.H. and Bates, C., Jr., Human engineering testing and malfunction data collection in weapon system test program. Wright-Patterson Air Force Base, Ohio: Wright Air Development Center Technical Report WADC TR 60-36, February 1960.

[15]Zeller, A.F., Human limitations and aircraft design. Air Force - Industry Conference, 1955.

but what the actual error rates are. If we know the error rates. we can plan around them or try to reduce them and evaluate the effectiveness of whichever course is taken.

We do have unpublished, classified data showing that mission failure due to human error is four times as frequent as that due to component failure in weapon drop tests. We also have a rough idea as to how the various species of human error are generally related to the total life cycle of equipment, and these are diagrammed in Figure 5-3.

The effects of assembly and installation errors. of course, tend to decrease with time as faulty units are detected and replaced in equipment checkout, maintenance, and retrofit programs. There is usually a slow startup of operations and some initial learning effect in both operator and maintenance errors; then, the operator error rates tend to stabilize, but maintenance errors tend to increase with increases of component failures during the wearout phase of components. This is a rough notion, but it may give you something to think about, for it has implications for the question: What are we predicting to? And it has some relevance to the meaning of error rate data collected at different phases of the life cycle.

Second only to the lack of field feedback data, the major problem in human error analysis is the variety and unevenness of the data available. Of necessity, we must often use data at its face value, but the data vary widely in terminology. manner of development, and level of reporting. Any efforts at standardization of these matters will greatly aid the progress of prediction techniques.

Along these lines, we prefer to call our figures "human error rates," because this is a straightforward, unequivocal, and generally acceptable concept; it describes exactly the kind of information we can use most effectively; and the acronym, HER, is guaranteed to get attention. More euphemistic terms such as "human reliability," "zero defects," or "human success probability" mean different things to other specialists, such as flight surgeons. quality inspectors, and personnel people.

Most people seem to be ready to accept the fact of human error, and this fact can be dealt with more effectively if dealt with openly. Too. if it is called "human error," it is more likely to be dealt with by behavioral scientists, as it should be. It is both useful and important. however. to distinguish, as Rook does, between situation-caused errors (SCE) and human-caused errors (HCE). Emphasis on SCE, especially when setting up error collection programs, helps remove the unfortunate and inappropriate onus attached to the words "human error."

Concluding Notes

SHERB. then, is a small file as yet; more an idea than an actuality. But it is growing, and it is a very useful and necessary adjunct to human error prediction. for the accuracy of such predictions and the effort required to develop them depend heavily upon the availability and accessibility of reasonably solid and generalizable data. upon the "knowns" of human performance.
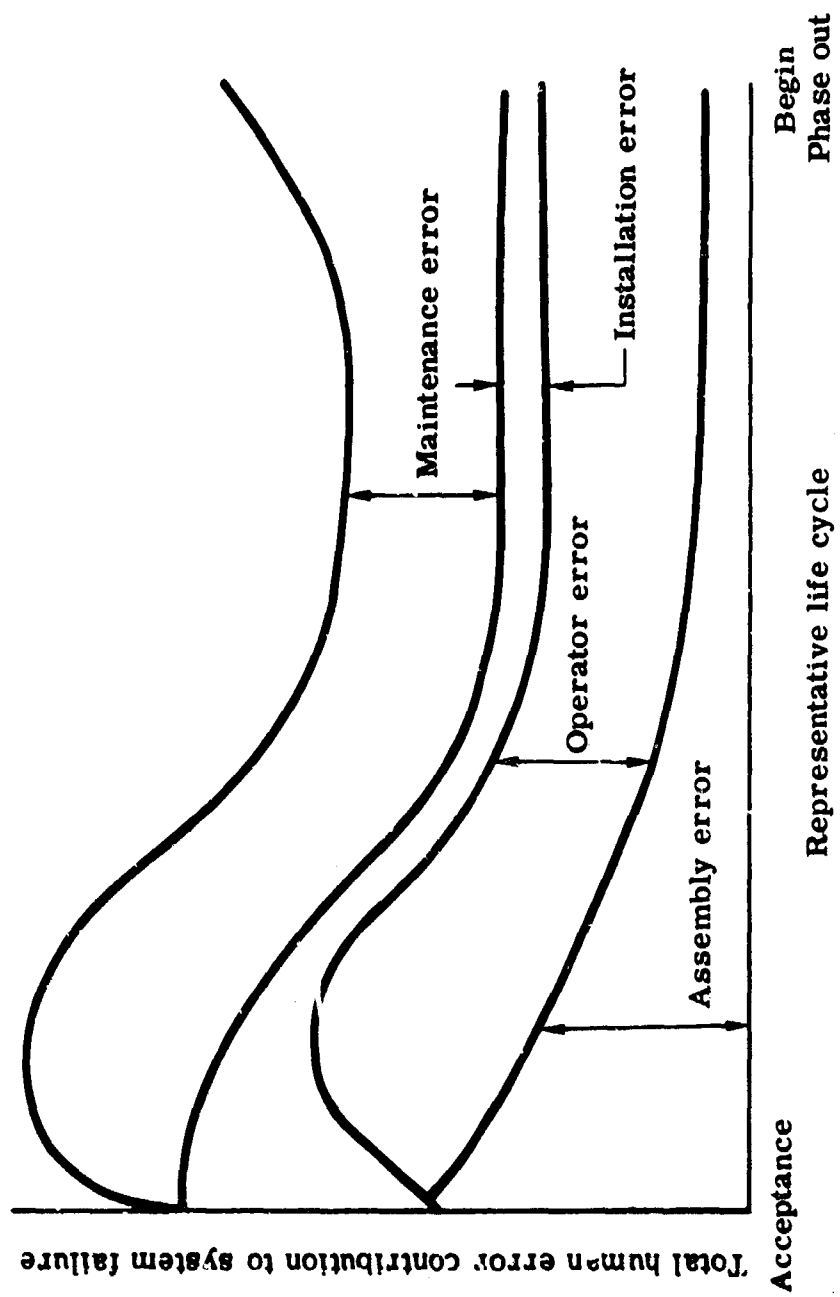
Figure 5-3. Proportional Contribution of the Different Species of Human Error to System Failure

When the file is more presentable, perhaps it can be published in full. In the meantime, we would be interested in exchanging such information with those of you who are developing comparable files of your own. And for those of you who are not developing such files, may we suggest that you consider it. You will be surprised at how useful it will become.

Obviously, the data currently available leave much to be desired. Merely complaining about this will accomplish little. Rather, it is the responsibility of every human factors specialist to specify what he needs, to determine how it should be collected, and to state clearly the value of having it. As soon as the human factors community acts in concert in this fashion, we will have good human error rate data; and there does not seem to be any aspect of human or man-machine performance that cannot be meaningfully interpreted in terms of human error.

6. A PRAGMATIC APPROACH TO THE PREDICTION OF OPERATIONAL PERFORMANCE

David Meister
Bunker-Ramo Corporation

The pragmatic approach referred to in the title of this paper assumes several things:

1. There is conscious attempt to avoid mathematical models and theoretical internal behavioral processes in developing predictions of operator performance. Of course, one cannot avoid these completely, but the goal is to extrapolate predictive indices directly from empirical data. These predictions assume that data can be generalized so that operator performance on equipment X can be used to predict operator performance on equipment Y if the two equipments and two operator populations are similar.

2. The emphasis in the pragmatic orientation is therefore on data, not theory. With this orientation data will be accepted from any source, even though these data may be less than completely precise or complete or otherwise tainted by inadequacies. The pragmatist will attempt to maximize whatever he has.

3. Nevertheless, he recognizes that his predictions must involve or be organized around certain parameters which are assumed to be important to operator performance; these will be discussed below. These parameters however, are selected primarily on the basis of h:s concept of "real life" equipment operations. This permits him to take full advantage of his logical and experimental prejudices.

It may appear from the above that the pragmatic orientation is overly simplistic, possibly even naive. In view, however, of the appalling lack of data to act as a foundation for theory construction, elaborate theories, particularly those possessing great mathematical sophistication, appear to be largely exercises in fantasy.

Despite this, anyone who is acquainted with the author's previous papers on the subject of predicting operator performance[1,2] is aware that there is considerable correspondence between his orientation and that of, for example, Altman[3], Blanchard[4], and Swain[5].

[1]Meister, D. Methods of Predicting Human Reliability in Man-Machine Systems. Human Factors, Vol. 6, No. 6, 1964.

[2]Meister, D. Development of Human Reliability Indices, in Proceedings, Symposium on Human Performance Quantification in System Effectiveness, Washington, D.C., January 1967.

[3]Munger, S.F. et al. Data Store: An Index of Electronic Equipment Operability. Report AIR-C43-1/62-RP(1), American Institute for Research, Pittsburgh, Pa., January 1962.

[4]Blanchard, R.E. et al. Development of a Technique for Establishing Personnel Performance Standards (TEPPS): Phase II Final Report, Dunlap and Associates, Inc., Santa Monica, California, January 1966.

[5]Swain, A.D. A Method for Performing a Human Factors Reliability Analysis. Report SCR-685, Sandia Corporation, Albuquerque, New Mexico, August 1963.

Like them, he finds it necessary to analyze system operations into relatively discrete units of behavior (e. g. tasks or sub-tasks) to which predictions can be applied. These units of behavior are organized around relatively molecular control-display components (e. g. knobs, dials, meters, toggle switches) which appear to be practical dimensions for describing the many different equipments for which one must predict. Like his colleagues also, the predictive data applied to these behavioral units are phrased probabilistically. The predictive indices applied are extrapolations of success/failure ratio data (i. e. s/n, where (s) is the number of successful completions of a task and (n) is the number of times that task has been attempted). The prediction is usually phrased in four figures, e. g. .8763. Since the sub-task or task unit level is relatively molecular, predictions for these units must be progressively combined to develop predictions for more molar units like system functions. This is done using a mathematical equation which describes the interrelated operation of these tasks as a guide. Hopefully one arrives at a single predictive value for the effectiveness of personnel operating the total equipment, subsystem or system.

As a pragmatist, one is concerned mainly about two things: (1) the data needed to make meaningful predictions of operator performance; (2) the ways in which the necessary data can be secured. These form the two themes of this paper.

It is a commonplace of meetings such as these to bewail the absence of sufficient data. As the author discovered in attempting to develop tables for predicting operator performance[6], there are some data, enough to make a start at prediction, but hardly enough to be satisfied with the predictive results. Since it is foolishness to contemplate the task of gathering all possible data on all possible parameters, the question arises; what data are needed. Until this question is answered, not much can be done to structure the data gathering process.

Both logically and heuristically it can be assumed that a restricted number of parameters account for the greatest part of the operator's performance. This assumption is a matter of faith as well as logic, because if human performance were equally affected by all possible factors, it would be infinitely variable and hence unpredictable.

---

[6]These tables were developed for the Rome Air Development Center under contract AF30(602)-4020. The purpose of this contract was to develop methods by which the Organization. Cost and Effectiveness characteristics (human performance prediction being included under Effectiveness) could be evaluated at the proposal stage.

These parameters tend to be task-oriented or at least to be related to operational system requirements. The significance of a parameter or its importance to performance will vary as a function of the conditions under which the parameters are exercised. Resolution is considered, for example, to be a significant parameter, as will be seen below, but only if the equipment being operated involves displays and only if these displays require perceptual functions which are significantly affected by poor resolution (an on-off light without a legend on it would be relatively immune to this parameter). If these conditions do not exist, resolution can be ignored. Those parameters whose effect, even when exercised, is minor, can presently be ignored for predictive purposes. As more empirical data are secured, these minor parameters can be incorporated in the prediction and predictive efficiency should increase.

The parameters selected as significant obviously define what data are needed, since the review of the literature performed in order to develop the predictive tables referred to previously revealed that no parameter is described by an adequate amount of data. There are obviously a host of possible parameters, as Altman[7] has pointed out. Some of the parameters finally selected (e.g. resolution) may be found in Altman's Data Store and represent rather fundamental (molecular) human engineering characteristics. Others, like the perceptual-motor or decision-making function performed by the operator, are relatively molar. The criterion used in selecting a parameter as important was: is it reasonable to expect in the operational situation that a major change in the value of the parameter will produce a major change in operator performance. Many of the human engineering characteristics about which experimental studies have been performed (e.g. the effect of toggle switch angle of throw) were rejected because their effect was considered trivial.

The unit of behavior for which one predicts is composed not only of the individual control-display component which is operated to perform a given function (e.g. tracking), but also the parameters which influence the operation of that component. One cannot, for example, predict the probability of successfully throwing a toggle switch unless one includes as factors in the prediction the number of other switches in which the one switch is embedded, the organization of these switches and the sequence of their activation. Hence, in order to develop predictive data it is necessary to specify not only the component but also the particular parametric conditions under which that component is being operated. The discussion below will describe what are considered significant parameters and the conditions under which relevant data can be secured.

---

[7]Altman. J.W. Classification of Human Error, presented at the meeting of the American Psychological Association. September, 1966.

Which parameters were selected as being important? As was indicated previously, the most elemental dimensions which appear to influence the operator's response are those which describe his controls and displays: (1) their number; (2) their organization; and (3) the sequence in which they must be utilized. The fact that these dimensions are so fundamental would lead one to anticipate that considerable information would be available concerning them; in fact, there is practically none.

Data must therefore be collected concerning the effect on performance of the number of identical components from which the control(s) to be activated or the display(s) to be read must be selected. Although at any particular moment in operation the operator responds to the single control or display to be utilized, that control or display is usually embedded in a number of other controls and displays. The selection process requires the operator to discriminate the single control or display from the surrounding others. Presumably the larger the number of embedding controls/ displays, the greater the difficulty of discrimination and the lower the probability of successful response. This parameter has been restricted for convenience to identical hardware components, because it is assumed that where controls and displays are recognizably different from each other, the problem of selection is much less. However, it would be desirable to determine the probability of correct utilization as a function of the total number of controls and displays, regardless of type, on a control panel.

It is relatively simple to determine by observation the number of identical or non-identical controls/displays which form the embedding context. However, this determination is influenced by the organization of these controls and displays.

This organization may be modular (side by side, horizontal or vertical) or non-modular (located in various positions around the control panel). What makes an organization modular is the tieing together in the same physical control panel area of functionally related controls and displays. If components are not organized into one panel area by some principle involving relationship among the components, they are considered non-modular. An exception may arise where the number of controls and displays in very few and/or are arranged strictly according to the operating procedure. In that event, the arrangement would also be considered modular. It is assumed that the probability of successful response is lowered where organization is non-modular.

Obviously, in any particular case a decision is required about what constitutes the module, but this judgment should not be too difficult. To determine the number and the organization of the controls and displays to which the operator must respond, one must first determine the responses required in any single procedural step (where the operating task involves more than one step). One must have or be able to develop at least a rough operating procedure which can be broken down into its component steps.

The third elemental parameter is the sequence of control-display
use. Sequence refers to the order in which controls and displays must
be used sequentially in successive operating steps or in which more
than one control must be activated or more than one display read in
the same procedural step. If that sequence of activation or reading
conforms to the arrangement of the controls and displays on the
equipment, it is called a fixed sequence. For example, if a module
contained four switches in a row and the operator had to throw them
in 1, 2, 3, 4 order within the same step or in a sequence of four
steps, the sequence is fixed. If, for some reason the operator had to
throw them in order 2, 4, 3, 1, the sequence would be variable. If
the switches were non-modularly organized, and if they had to be thrown
in an order which bore no relationship to their location, this would be
considered also a variable sequence. It is assumed that the probability
of successful performance is less when the sequence is variable.
Sequence can be determined by observation of equipment operations or
by analysis of an operating procedure.

One should also know something about the accuracy required of the
operator in performing a task. This accuracy may be of two types:
(a) determined by the nature of the control-display component or (b)
by an operational requirement which sets a criterion of successful
performance. The first type is exemplified by a scale on a complex
meter which requires interpolation; the second by an operational
requirement that no more than two errors be made in inputting any message
sequence. One would assume (this is only an assumption because precise
data are lacking) that a quantitative meter demands more accuracy of
the operator than does a qualitative one; typing a rough draft requires
less accuracy of a typist than does typing a final draft. Presumably
the probability of successful response is lower when required accuracy
is greater.

This kind of information too should be fairly easy to determine by
analysis of the control-display component and the operating procedure.

Another parameter for which information is needed is what we
call operator loading or pacing. The essential element in this
parameter is that the operator must respond at some rate other than
that which he would ordinarily assign to himself. Where the operator
himself controls the speed with which he responds, loading is absent.
Where the operator must respond as rapidly as he can (i.e. with
some strain) or at a speed which must match the rate at which stimuli
are presented to him (provided that rate is faster than his normal
response speed), he is considered to be loaded. It is assumed that
the probability of successful performance is less when the operator is
loaded.

This type of loading is peculiar to time stress and is not
assumed to represent a generalized "anxiety" state. Obviously
operator loading varies on a continuum, but in terms of the tables
referred to earlier loading has been arbitrarily conceived as a binary
factor, i.e. it is or is not a significant factor. The reason is that
we have very little data on the effect of different amounts of time
stress on performance. Time stress can sometimes be inferred from
the operating procedure or by observation of operator performance
(including interviews) in the operational environment. However, for
precise data, quantitatively relating time stress to performance
success, experimentation is required.

Display exposure time also varies infinitely. Moreover, the criterion of what is adequate exposure time (from an operator performance standpoint) depends to a large extent on what must be discriminated and the context of that discrimination. Since data dealing with the effect of exposure time on various functions are largely lacking, it has been necessary in constructing the predictive tables referred to earlier to assume (based on the very little data available) that any exposure time less than 10 seconds for a complex display is restricted, and to collapse the parameter into a binary condition: adequate and restricted. It is assumed there is a lower probability of successful performance in reading a display when its exposure time is restricted.

This parameter is one which it would be difficult, lacking proper controls, to study in the operational environment.

Display visibility may be acceptable or low, depending on whether or not the display meets standards of resolution, contrast or image distortion. If the latter are significantly below standards required for perceiving the display without strain, visibility is low. Presumably the probability of successful performance is lower under conditions of low visibility, taking into account the accuracy required of the task. However, again the amount of available data is quite small.

Actually, most systems are constructed with the proper display visibility and there is some suggestion in the literature that the effect of non-optimal visibility on operator performance is relatively small except for certain special complex display subsystems (e. g. radar) and tasks (photointerpretation). Like display exposure time it would be extremely difficult to secure precise data on the effects of display visibility in the operational environment.

The nature of the stimulus presented in a display (i. e. whether it is structured or unstructured) is also important to the operator's performance. A structured stimulus is one to which the operator responds directly and immediately, in terms of an already learned meaning (e. g. an alphanumeric character). In contrast, an unstructured stimulus must be analyzed in terms of its basic dimensions, before its meaning can be identified. For example, a sonar pip (which is unstructured) must first be analyzed in terms of size, shape and brightness before it can be categorized as a submarine. It is assumed that success probability is lower in responding to unstructured stimuli.

The number of visual stimuli displayed may vary greatly, of course, from a single alphanumeric on a CRT to massed columns and rows of alphanumerics on a large screen display. It is assumed that the probability of success in detection, discrimination and identification decreases as a direct (although probably non-linear) function of the number of visual stimuli the operator must respond to.

Although it is simple to determine the type of stimulus being presented, it is difficult to secure precise data on the effect of number of stimuli in other than a controlled, experimental situation. The specification of the type of hardware display often (but not always) indicates the type of stimulus presented by the display (e. g. a radar display usually -- but not always -- indicates an unstructured stimulus). Where this is the case it is unnecessary to apply a special predictive index (i. e. a standard error rate or failure probability) for this parameter, although one must consider it in developing predictive indices and in selecting a particular index for prediction.

Where the number of stimuli is determined by external systems (e. g. aircraft) it may be difficult to apply a standard predictive index to this parameter because that number is not a fixed quantity.

Operator <u>function</u>, defined in terms of the type of response specifically required of the operator by the task, is another crucial parameter. The functions involved are:

a.   discrete control response;

b.   continuous control response;

c.   simple monitoring (no detection required);

d.   detection;

e.   discrimination;

f.   perceptual-motor coordination (e. g. tracking);

g.   stimulus identification;

h.   information extraction (e. g. counting or updating stimuli);

i.   decision-making based on the coordination of information from multiple display sources.

This listing is of course not exhaustive and others might suggest variations.

While no linear continuum of difficulty can be associated with these categories, it can be assumed that, all other things being equal, success probability is greater with simple functions (e g. discrete control responses) than it is with more complex ones (e. g. stimulus identification or decision making).

It is relatively easy to determine the existence of simple functions, like control functions, because these are usually implicit in the control or display component (e. g. switch activation requires a discrete control response). For these functions special predictive indices are not required because they are implicit in the operation of the component. It is much more difficult, however, to identify the functions involved in operating complex equipment. At the moment it is unclear whether for these complex functions special predictive indices will be required, or whether they can be subsumed in the particular equipment (e. g. to assume that large screen displays always require discrimination and stimulus identification). Much more data will be required to answer this question.

Stimulus movement, as a parameter to be included in prediction of operator performance, is important only when the display involves moving stimuli. Obviously, that movement can vary over a range of values; hence the determination of performance data relative to this parameter can only be precisely gathered in an experimental environment.

Obviously controls and displays are activated not only separately but also (and probably more often) in a coordinated manner. Multiplying the performance probability for a control (e.g. .9843) with that of a display operated in coordination with the control (e.g. .8772) will not necessarily give one the same performance prediction (.8634) one would get if data are collected relative to the integrated operation of the two. Hence it is necessary to consider the characteristics of control-display coordination. This parameter is defined as activation of a control in conjunction with or in response to a display or perception of a display in response to a control activation. It may have the following variations:

a. activation of control is primary, the display being only feedback;

b. activation of a control to elicit a display reading;

c. activation of a control to adjust or match a display reading;

d. activation of a control in response to a display, which may include

   (1) activation as a response to a simple display pattern involving recognition of the onset of that display pattern (e.g. push the button when the light comes on);

   (2) activation as a response to complex display patterns involving discrimination of alternative display patterns or activation in response to information coordinated from multiple displays (e.g. perform response X when displays A and B occur together, but not if A or B alone occur).

These control-display relationships can be observed operationally, but their quantitative measurement (particularly the more complex relationships) will require an experimental environment.

It is also necessary to take account of the fact that more than one task may be performed concurrently by the same operator. Hence it is necessary to consider in one's predictions concurrent activities. It is assumed that where the operator must perform concurrent (although perhaps subordinate) functions at the same time he is operating his controls and displays, the probability of successful response is decreased. Among the major concurrent activities may be communicating information directly or via intercom, recording data, plotting graphs or other charts, filing, etc.

The operation of this parameter can be relatively easily observed. To secure data on the impact of this parameter, however. it will be necessary to compare two concurrent activities with the performance of each one separately. This can be done in the operational environment, but it will require careful selection of different operational situations.

The amount of information which the operator must handle obviously influences his performance. The definition of this parameter is extremely difficult, however, where complex control-display equipment is involved and it is unlikely that precise information about its effect can be secured except in the experimental situation. For present predictive purposes amount of information has been defined only comparatively, in terms of the number of categories of data presented in any one display channel. For example, a discrete indicator might present only two levels of information (e. g. power off-on); a qualitative meter might display three levels (bands) of information (e. g. in-tolerance, warning, and out-of-tolerance). It is probable that the greater the amount of information the operator must handle, the lower the probability of successful task completion.

A parameter which was considered, but which was not included in the predictive tables because of lack of data, is feedback. Feedback may be of two types: (a) direct, that provided directly by a display specifically designed to provide this information; (b) indirect, that provided by the progression of displayed equipment events or status which accords (or does not accord) with the operator's learned expectations of how the equipment should perform under normal conditions. Indirect feedback is always present in equipment operations, but because it is so nebulous, so difficult to define, it is not considered as one of the effective parameters. However, the provision of direct feedback should improve the probability of responding successfully. Direct feedback should be easily identifiable in the operational environment.

Data must be secured in terms of individual control-display components as influenced by the parameters assumed to affect the operation of these components. Table 6-1 indicates the parameters assumed to be effective (under specified conditions) in the operation of particular controls and displays. The control-display component is listed vertically, the parameters horizontally. An X in the matrix cell indicates that a particular parameter should be considered in determining the predictive value for a given control-display component.

It was indicated earlier that although many parameters may influence the operation of a given control-display component, not all of these are equally influential. (This is why it is possible to ignore some of them in developing the predictive indices.) Hence the large number of parametric interactions implied in Table 6-1 should not be too upsetting; in any given operational condition the predictor may decide, using his knowledge of that condition, to eliminate one or more of these parameters as being <u>in this condition</u> not important enough to warrant including.

A parameter was related to a control-display component in Table 6-1 if it was considered to have a potential effect, however slight. Certain parameters (i.e. operator function, concurrent activities) are related to all components, since each requires some behavioral function or could have another concurrent activity associated with it. In general, a parametric effect was singled out for attention in Table 6-1 if the predictor should consider the parameter in determining the predictive index. After examination, the parameter may be rejected as not being applicable to a given operation. For example, one must determine in all cases whether or not a concurrent activity is going on, but many cases will have no concurrent activity and one then simply ignores the parameter.

Every operating situation is obviously influenced by more than one parameter which exert their effect, not individually, but in interaction. For that reason, although it simplifies the predictive situation considerably, one can only artificially attach to the parameter a standard decremental value (i.e. error or failure rate) representing the influence of that parameter. (These parameters have a negative influence on performance because they complicate the operator's task and thus reduce the reliability of his performance, just as an additional component tends to reduce equipment reliability.) They have no positive effect (i.e. to improve the probability of successful performance, because the optimal situation is one in which the effect of the parameter is nil, that is, the parameter does not exist). So, for example, the absence of feedback in control activation might represent a penalty (error rate) of .0030 (invented number, of course) to be subtracted from the prediction of optimal performance (1.00). Nevertheless, Altman in his data store established standard performance probabilities for particular parameters; and it was found necessary to do the same in the predictive tables referred to earlier, solely as a means of simplifying the problem of handling the large number of interactive parameters.

How these parameters combine in relation to the same control-display component (i.e. additively or multiplicatively) is another problem which will be solved only when there is a sufficient amount of data available so that one can compare the effect of various parametric combinations.

How can one secure data on these parametric conditions?

TABLE 6-1. The Influence of Selected Parameters on Control-Display Utilization

Control-Display Component	Number of Components	Organization	Sequence	Required Accuracy	Operator Loading	Display Exposure Time	Display Visibility	Stimulus Structure	Number of Stimuli	Operator Function	Stimulus Movement	Control-Display Coordination	Concurrent Activities	Amount of Information	Feedback
Discrete Control (Switch, Pushbutton, Keyboard)	X	X	X		X					X			X		X
Discrete Control (Typewriter)	X	X		X	X					X			X		X
Continuous Control (Rotary Switch, Knob, Joystick)	X	X	X	X	X					X			X		
Discrete Indicator (Legend/Non-Legend Lights)	X	X	X		X	X	X			X			X	X	
Continuous Displays (Meters)	X	X	X	X	X	X	X			X			X	X	
Dynamic Displays (CRT/Large Screen Display)	X	X	X	X	X	X	X	X	X	X	X	X	X		X
Control-Display Combination (Discrete Control-Discrete Display)	X	X	X		X	X	X			X			X		
Control Display Combination (Discrete Control-Dynamic Display)	X	X	X	X	X	X	X	X	X	X	X	X	X		X
Control-Display Combination (Continuous Controls-Continuous Displays)	X	X	X	X	X	X	X			X		X	X	X	X
Control-Display Combination (Continuous Controls-Dynamic Display)	X	X	X	X	X	X	X	X	X	X	X	X	X		X

The experimental method of securing data is so familiar that it need only be referred to. Moreover, if there need be any other reason for de-emphasizing the experimental method in this paper, it is because the author confesses to a lingering doubt that the experimental work performed in the future will supply the necessary data. The reason is that the choice of the parameters to study and the means of studying them have been left largely to personnel with an academic orientation which is not responsive to the needs of the human factors discipline. In view of the extremely poor record which human factors research has to date in supplying requisite data (to develop his data store Altman found only 164 relevant studies, most of which are considered by this author to be irrelevant), one can hardly hope that the experimental situation will change very soon.

A major complaint against traditional human factors experimentation is that most of the tasks and equipment used are remote from the tasks and equipment used operationally; hence, the results are non-applicable. In addition, the experimental situation is complicated by the fact that while prediction is concerned with the successful/ unsuccessful performance of tasks, the experimental studies performed have emphasized errors and/or response time. In fact, in many cases the operation studied has not been an operationally meaningful task at all but rather an action with meaning only for the study. In other cases (too many, unfortunately) the raw data are not reported. Then again, in many studies only a comparatively few trials have been given (only enough to establish the base for some statistical test of confidence) so that the subject cannot be considered to be properly trained in his activity. Finally, a major limitation of experimental studies has been the use of a non-applicable subject population (usually college students).

If one cannot rely on experimentation to provide the requisite data, why not attempt to gather what one can from the operational situation?

The problem of data collection in the operational environment is not one of measurement per se, since the measurement of task success/ failure requires merely counting of the frequencies of such successes or failures. (Task success/failure is a purely binary condition determined by the success criterion.) The difficulty in operational measurement is the setting up of conditions which permit one to isolate the parameters whose relationship to task success one is interested in. If one can identify the effective parameters in the operational environment, the measurement problem disappears. However, since parameters usually exist in interaction, it is almost impossible to isolate a single parameter.

The solution of the problem is to look for these parametric conditions in the operational environment which display the combinations of parameters one wishes to measure. Since the performance data one secures is always related to two or more parameters, it is necessary to find different combinations of these parameters in the operational situation, to measure them, and then compare the results. Thus, one might look for an equipment or system which involved few unstructured stimuli and compare it with a similar situation involving few structured stimuli. Differences would suggest the effect on performance of types of stimulus.

Thus, this author feels that despite the manifest difficulties
in operational data collection, some useful data can be gathered.
In that way one would be so much the more ahead of the game; in
addition, collecting data on those parameters which can be collected
in the field might act as a spur to the experimentalists by showing
them what can and should be done, even under non-optimal data
collection conditions.

To gather data operationally a pragmatic strategy is suggested:

(1) On the basis of the components and parameters listed in
     Table 1, determine which equipments and parameters one
     wishes to collect data on and examine the available
     operational situations for the one(s) most closely resembling
     those desired. This examination would involve not only an
     analysis of the equipment's control-display components, but
     a review of its operating procedure. This is necessary
     because where an equipment includes in its operation (as
     many do) a number of different control-display components
     and tasks, the operation must be broken down into the
     sub-tasks which pertain to these control-display components.
     It is necessary also to determine how the sub-tasks are
     related to the overall operating goal; this in order to specify
     the criterion of successful task completion.

(2) Describe all of the major parameters which can be isolated
     by observation of the operational situation. This is necessary
     if one wishes later to compare this operational situation
     with others. Collect data by performing the necessary
     measuring operations.

(3) Repeat this process for other operational situations involving
     the same equipment operation with different combination of
     parameters (e. g. structured vs. unstructured stimuli) or
     different parametric values (e. g. restricted vs. adequate
     exposure time). Collect data on these other situations.

(4) Compare the results of studies involving the same equipment
     components but different parametric conditions. If a sufficient
     number of parametric conditions have been sampled, it will be
     possible to assign differences in performance to differences
     in these conditions. Thus, if the only difference between sets
     of parametric conditions is one of resolution, then a particular
     decremental value can be assigned to the resolution parameter
     In a very few cases it was possible in developing the tables
     of predictive values referred to earlier to make such a
     comparison (very tentatively, of course). Where comparisons
     are confounded (e. g. two operational situations contain the
     following parametric combinations: (1) modular organization,
     adequate visibility low required accuracy; (2) modular
     organization, restricted visibility, high required accuracy) it
     will be necessary to estimate the relative contribution of the
     visibility and accuracy parameters to the difference in
     performance found in the two situations. If there are no
     clues in the operational situation, an answer might be to divide
     the performance variance in half and assign each half to each
     parameter. This is a calculated risk which will provide at the
     least an approximation of the correct values. Sampling
     additional operational situations should progressively provide
     more valid data.

With regard to the operations of gathering the desired data, there appear to be two ways of proceeding: have the operators themselves report; or send out special teams (probably of engineering psychologists) to observe. There are reasons (too lengthy to go into this paper) why the latter alternative is preferable. If the latter method is used, the observer must learn the details of equipment operation before he can observe; but this is an acceptable penalty.

In summary, what is required is a consistent, long term effort to secure predictive data. It is doubtful whether the experimental milieu will provide the necessary data; so attention must be paid to gathering these data operationally. Is this possible? Will the human factors establishment support it? It will be interesting to see what happens in the future.

7. MAN MACHINE SIMULATION – THE PIMO APPLICATION

Glenn Spencer
Serendipity Associates

Serendipity Associates is currently under contract to the Air Force to develop a new approach to the presentation of the technical data used by maintenance technicians, otherwise known as T.O. s. This project, termed PIMO (Presentation of Information for Maintenance and Operation), has been under way for almost two years and is currently in the test and evaluation phase. As depicted in Figure 7-1, the project has resulted in the development of an audio-visual approach to on-aircraft and in-shop maintenance information presentation. The test and evaluation phase is devoted to establishing the actual effectiveness of this system in the operational environment. In addition, the differential effectiveness of audio-visual and booklet presentation is being evaluated. The purpose of this paper is to discuss the approach and means used in the effectiveness evaluation effort. Specific attention will be paid to the digital simulation model which was employed and the types of maintenance variables of concern to the study.

System effectiveness analysis has played a major role throughout the execution of project PIMO. The primary objective of the effectiveness effort is to establish the advisability of investing in a system which would improve the manner in which technical data are presented to maintenance technicians. Also, effectiveness data are used to aid in the decision as to which of a set of alternative systems should be adopted, given the cost and expected benefits of each.

The object system for the current test phase is the C-141A jet cargo aircraft operated by the Military Airlift Command. The C-141A is rapidly becoming the backbone of MAC's airlift fleet and has contributed greatly to the excellent logistics support of U.S. forces in South Vietnam. The system is not without its problems, however, as indicated by the increasing rate of mission delays due to maintenance (from 4% in January 1966, increasing to 13% in January 1967).

Some time could be spent describing the history of project PIMO and the maintenance problems of the C-141A and this would aid in the understanding of the role of the simulation model; however, time constraints require that these preliminaries be skipped in order to enter immediately into the discussion of the specific means used in the effectiveness analysis.

It was recognized early in the project that in order to make the benefits of information presentation improvements meaningful, they had to be expressed in terms of the object system, namely the C-141A. The conceptual basis for this decision is that the value of the requirements maintenance system is derived from object system requirements and, thus, changes in performance at the support level must be evaluated in terms of object system performance and/or cost. Since the objective of project PIMO is to improve tech data presentation, the immediate impact will be on the performance of maintenance technicians. Therefore, some means had to be devised which would relate changes in maintenance performance to changes in C-141A effectiveness. As will be discussed later, the means employed was the AMES (Aircraft Maintenance and Effectiveness Simulation) model. To better understand how the AMES model relates maintenance performance to system performance
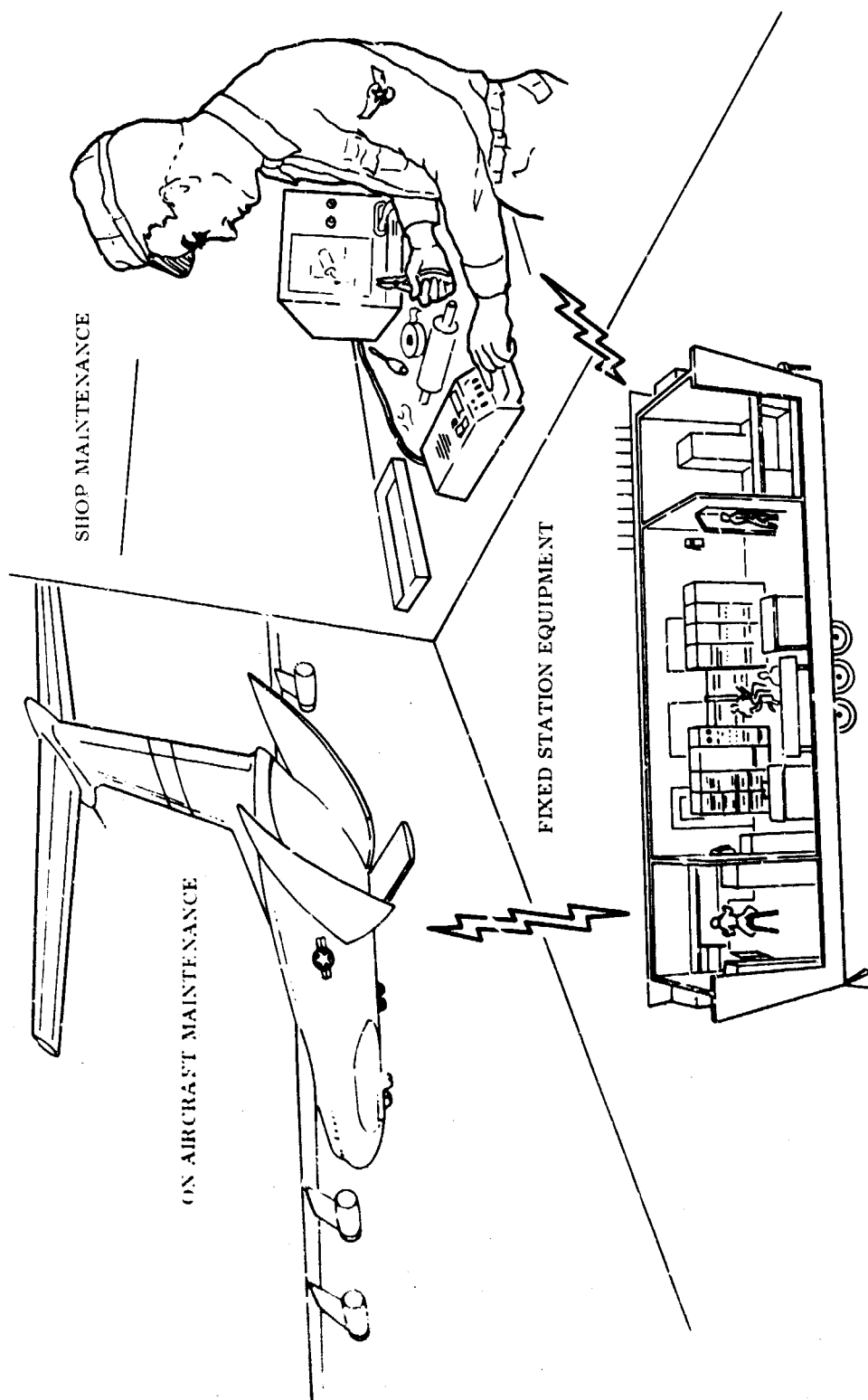
SHOP MAINTENANCE

ON AIRCRAFT MAINTENANCE

FIXED STATION EQUIPMENT

Figure 7-1.   Presentation of Information for
Maintenance Operation

7-2

it will be necessary to devote some time to a description of the types of maintenance performance measures used.

Maintenance Performance Measures

During the past five years Serendipity Associates has developed and refined a method of expressing maintenance function performance in terms which are (1) measurable, using existing data, and (2) relatable to higher system objectives. The approach is based on the concept that the outcome of maintenance functions can be expressed in specific "state" terms and that the performance of the maintenance function is characterized by the resources and time required to achieve the output state and the likelihood that the output state is correct.

The overall objective of the maintenance support system is to change the state of a system from one of "malfunctioned" to one of "functioning properly". Individual maintenance functions can be separated into two classes; informational change of state functions such as trouble-shooting, pre-flight, and operational checks and physical change of state functions such as remove/replace and calibration.

The objective of an informational change of state function is to determine whether or not a system, subsystem or component is go or no-go and, if no-go, which item is causing the no-go state. From the standpoint of functional reliability it is possible to make the following types of errors in an informational change of state function.

TYPE I   -   Erroneously designating the state of a system as bad (good called bad).

TYPE II   -   Erroneously designating the state of a system as good (bad called good).

TYPE III -   Erroneously identifying the source of a malfunction. (Wrong part isolated).

Errors of the above type can and do happen in the performance of a maintenance function. In certain cases errors are made, discovered and corrected during the function and the only effect on the output state is one of performance delay. For this reason another type of error is defined as follows:

Type t error - Delaying the execution of a function beyond some inherent performance time.

In physical charge of state functions it is possible to damage parts in installation or removal or during calibration or adjustment. For these functions the following error type is defined.

TYPE d error - Damaging or otherwise incapacitating a system during the process of repair

It is important to note that the above are measures of functional reliability and do not necessarily imply human error. A bad system could be passed during pre-flight merely because the procedures do not call for a check. Many things can affect functional reliability including test equipment, training, technical data and procedures, to name just a few. It is the case that these types of errors are oftentimes viewed as human errors, however, for the purpose of the PIMO study, strict adherence was made to the use of the term functional reliability

Based on the above definitions of functional reliability an approach was devised to indirectly measure the probability of occurrence of each type of error. When dealing with the C-141A a system functional flow-logic diagram is used to depict the basic operational and maintenance functions. (Figure 7-2). The output states of each function are identified as are the flow of aircraft, parts, and information.

As shown on Figure 7-2 each function is identified with the types of error which can occur within it. By gathering data on the performance time and output state of each function it is possible to estimate, indirectly, the above mentioned event probabilities (The term indirect is used to differentiate the evaluation from those which depend on direct observation.) Actually the best way to describe this approach is to give an example.

Suppose that aircraft COO1 has, according to the navigator, experienced a malfunction in the search radar system and that the symptom was "faulty video". Following the maintenance actions on this aircraft, the maintenance personnel reported "checked o.k., no maintenance required", and the aircraft was allowed to depart on its next mission (new crew). At the next stop the navigator station again reported trouble with the search radar, again "faulty video". This time, however, the ground crew isolated the problem to the receiver which was replaced. The removed item was bench checked and repaired and returned to base supply. No problems were reported against the search radar for six subsequent mission legs and 50 flying hours.

The above sequence of events leads to the deduction that a type II error (accepting a bad system) was committed in the troubleshooting function following the original complaint by the aircrew. The factors to be considered are:

1. The repeat of the complaint against the radar on the next flight leg with the same sympton description.

2. The requirement for repair of the receiver which was removed.

3. The absence of repeated squawks against the radar set.

Information on detailed maintenance actions such as that shown above can be used to indirectly compute the other functional error event probabilities. Time does not allow a complete description of the technique, however, it should be pointed out that the approach depends on the acquisition and correlation of a variety of maintenance and system operational data as well as the considered opinion of knowledgeable engineering personnel. It has been found, however, that such data are available if one spends enough time researching the sources. The primary date sources used in the C-141A maintenance reliability effort were:

AFM 66-1 AFTO Forms 210/211

Form 99-2 Specialist Dispatch Records

MAC Form F-1 Mission Following Records (MAC Specific)

MAC Base - Specific Aircraft Status Sheets (MAC Specific)

Figure 7-2. AMES Functional Flow Diagram

Form 781 - Aircraft Log.

The data obtained from these forms allowed analysts to follow the history of individual aircraft maintenance actions including the shop actions taken subsequent to component removal from a specific aircraft.

Since these reporting systems also contain overlapping information it is oftentimes possible to correct erroneous entries and to fill in voids by properly cross-checking reports. This procedure adds significantly to the reliability of the data base.

The analysis procedure is basically one of flagging "maintenance repeats" on the same system or a functionally related system for subsequent analysis by personnel familiar with the system design or functional characteristics. Candidate errors are analyzed in light of overall system failure rate and an analysis of shop actions on removed components, if any occurred.

A summary of the maintenance function reliability analysis is shown in Figure 7-3. The error rates shown on this chart were obtained from a detailed analysis of all maintenance records of approximately eight C-141A aircraft over a period of six months. Statistical tests indicate that the sample size used was adequate.

Two independent studies of C-141A reliability, one performed the Aviation Week and Space Technology magazine, and another by a Operational Evaluation Group for the C-141A at Travis AFB California, have tended to substantiate our findings. In the Aviation Week article (AW Feb 13, 1967, pg. 30) data were presented which showed significant differences in the reliability of similar items when used on the C-141A and when used by the airlines. Although not substantiated, it was the opinion of the author that the major contributor to this difference was the skill level of attending personnel.

Of greater immediate significance was the finding by the Operational Evaluation Group (OEG) that an average of 30% of all C-141A components received by the avionics shops for repair are checked o. k. This figure is consistent with the probability of a type III error (erroneous fault isolation) computed by Serendipity. The effects of erroneous component removal are considerate in light of the pipeline time involved in spare parts logistics.

Refinements are currently being made to the error analysis procedures to reduce the level of judgment required, however, this element cannot be eliminated entirely. Serendipity Associates is convinced, however, that the overall approach is sound and provides the measures necessary to link maintenance effectiveness and system effectiveness.

PIMO Field Tests

The data analysis effort was aimed at identifying a base line from which the effects of the PIMO system would be measured. In addition to the identification of functional reliability, data were obtained on function performance time, function resource requirements and personnel assignment policies.

The effect of alternative approaches to PIMO on maintenance performance was measured through the means of a field test wherein performance time and reliability were determined by comparing personnel performance with the current T O. approach to information

ERROR RATES

SYSTEM	I - PF	I - M	I - PO	II - TR	III - TR	d - CA	II - VR
Airframe	0.0%	0.0%	0.0%	0.0%	1.1%	9.6%	70.0%
Cockpit and Fuselage Compartments	0.0%	0.0%	0.0%	0.0%	0.0%	6.7%	50.0%
Landing Gear	0.0%	12.5%	0.0%	0.0%	3.1%	14.6%	45.5%
Flight Controls	2.3%	13.8%	0.0%	9.4%	10.4%	6.5%	23.5%
Turbofan Power Plant	0.0%	19.6%	0.0%	4.8%	4.9%	7.2%	71.4%
APU	0.0%	3.6%	0.0%	6.2%	0.0%	0.0%	--
Air Conditioning, Pressurization, Etc.	0.0%	35.7%	0.0%	7.7%	2.6%	3.4%	25.0%
Electrical Power Supply	0.0%	7.1%	0.0%	0.9%	11.1%	4.4%	50.0%
Lighting System	0.0%	5.4%	0.0%	1.8%	0.6%	5.7%	80.0%
Hydraulic Power Supply	3.6%	7.1%	0.0%	11.1%	3.7%	9.5%	66.7%
Fuel System	0.0%	14.3%	0.0%	8.3%	0.0%	6.2%	0.0%
Oxygen System	0.0%	8.9%	0.0%	5.9%	5.9%	8.3%	50.0%
Misc. Utilities	0.0%	5.4%	0.0%	4.5%	4.5%	4.9%	75.0%
Instruments	3.6%	32.2%	1.8%	4.3%	7.1%	4.2%	53.5%
Autopilot	0.0%	12.5%	0.0%	6.5%	8.7%	0.0	0.0%
HF Communications	0.0%	0.0%	0.0%	0.0%	12.5%	12.5%	50.0%
VHF Communications	0.0%	5.4%	0.0%	11.8%	17.6%	0.0%	66.7%
UHF Communications	0.0%	1.8%	0.0%	13.8%	6.9%	10.7%	60.0%
Interphone	0.0%	1.8%	0.0%	10.0%	7.5%	12.8%	82.5%
Radio Navigation	0.0%	1.4%	0.0%	3.1%	30.6%	2.1%	50.0%
Radar Navigation	0.0%	8.8%	0.0%	6.7%	33.0%	4.5%	65.5%
OVERALL	9.5%	197.3%	1.8%	4.5%	8.7%	6.4%	57.7%

Figure 7-3. C 141A Maintenance Reliability

presentation, a reformatted booklet approach and an audio-visual approach. The results of these tests pointed strongly to the audio-visual approach; however, the differential performance measurements were insufficient in and of themselves to justify a major modification of the T.O. systems. Therefore, the AMES model was employed to express these performance improvements in system terms.

The AMES Model

The AMES model is a digital simulation model programmed in FORTRAN IV for the IBM 7094. The model is basically a representation of the system diagram shown in Figure 7-2. Each of the functions of this diagram are represented by subroutines which determine if the function can be initiated using available resources, determines the time required to perform the function and simulates the effect of errors. Supervisory routines control the movement of aircraft, parts and resources such as personnel and manage system inputs such as mission demands. The basic structure of the model is logically consistent with the functional approach used in the maintenance reliability analysis.

A complete squadron of 20 aircraft can be handled simultaneously including up to 20 maintenance actions per aircraft. Removed components are traced through shop or depot repair and into base supply. Bad components resulting from erroneous maintenance enter Base Supply and affect the probability of removing a bad item for installation on an aircraft.

Maintenance function errors are simulated by maintaining two states for aircraft and components; the actual state and the apparent state. Inherent failures can occur only in the Mission function or as the result of a damage error ("d" error) in the Repair function. A failure or damage error establishes that the actual state of a system and component is bad. It is the duty of air crew and/or ground crew personnel to identify the apparent state of the system. If this is done error-free the actual state and the apparent state will be identical.

For example, assume that a failure occurs in the Radio Navigation system during flight and is correctly reported by the aircrew. Upon entry to the Troubleshoot function the actual and apparent state of the system are "bad", and a random number is drawn and compared to the probability of a Type II error occurring. If such an error occurs, the apparent state of the system is set to "good" and the system is allowed to remain actually bad, apparently good. If the Preflight function does not adequately check the Radio Navigation system it is likely that the failure will remain in the aircraft through the next flight, possibly causing abort depending on the probability factor entered for this system.

By following aircraft and aircraft components in this way it is possible to relate changes in maintenance function performance time and reliability to system performance effectiveness. An error in trouble-shooting may cause a mission abort and will ultimately require a follow-up maintenance action and thus additional aircraft ground time.

Maintenance function variables have an indirect effect on the object system in that they interact with factors such as resource availability and personnel utilization. As function time is reduced fewer delays are incurred due to the lack of personnel and equipment since these resources are idle more frequently. In addition, improvements in function reliability reduces the demand for spare components which are erroneously removed (type III errors), thereby reducing delays for lack of these parts. By accounting for spare components, individual items of AGE

and the availability of personnel, the AMES model simulates these interactions so that the true effect of maintenance performance improvements is measured.

The personnel required to perform a given function on a given system are specified by model input data. Up to twenty different types and/or skill levels can be identified. Delays for a specific type of personnel, say skill level 5 radar technician, can occur even though less experienced personnel (level 3) of the same type are available, as long as the assignment policy calls for level 5 only.

The AMES model is designed to allow the analyst to study alternative personnel assignment schemes by providing a "secondary" personnel set which is to be used in the event that the "primary" or preferred set is unavailable or by changing the original data set from run to run to reflect different policies. Thus, the effect of improved information presentation may be reflected in terms of increased utilization of lower experience level personnel and, in turn, increased system effectiveness by increasing personnel availability.

The primary C-141A system effectiveness measure of concern is flying hours per aircraft month, or aircraft utilization. System availability is usually used as the measure of maintenance system effectiveness, however, aircraft utilization, while more difficult to compute, provides a direct entry into cost effectiveness analysis since the Military Airlift Command (MAC) costing system is based on ton miles of airlift capability. Changes in aircraft flying time can be converted directly into ton miles and then into value-added. The underlying assumption is that mission demands exceed aircraft availability and this has been verified through MAC headquarters.

In addition to aircraft utilization the model provides other measures of C-141A effectiveness such as departure delay time and mission cancellations rate to assure that aircraft utilization is not gained at the sacrifice of other important considerations. These measures are not, however, readily expressible in cost terms.

Simulation model runs are made using existing system data to provide a baseline effectiveness level. Basic maintenance function input data (time, reliability and personnel skill level requirements) are then changed to reflect improved performance and the simulation is re-run. This procedure is followed until a parametric relationship is established between the maintenance variables and system effectiveness.

Figure 7-4 represents the results of the parametric analysis for the C-141A using the AMES model. This graph shows the relationship between time in function (TIF), error rates, personnel assignment policies and percent increase in flying hours. The personnel assignment policies were as follows:

POLICY 1 - Personnel assigned in accordance with present policies i.e., lower experience levels not allowed to operate independently.

POLICY 2 - Lower experience levels allowed to perform repair and test functions if middle skill levels unavailable - no troubleshooting.

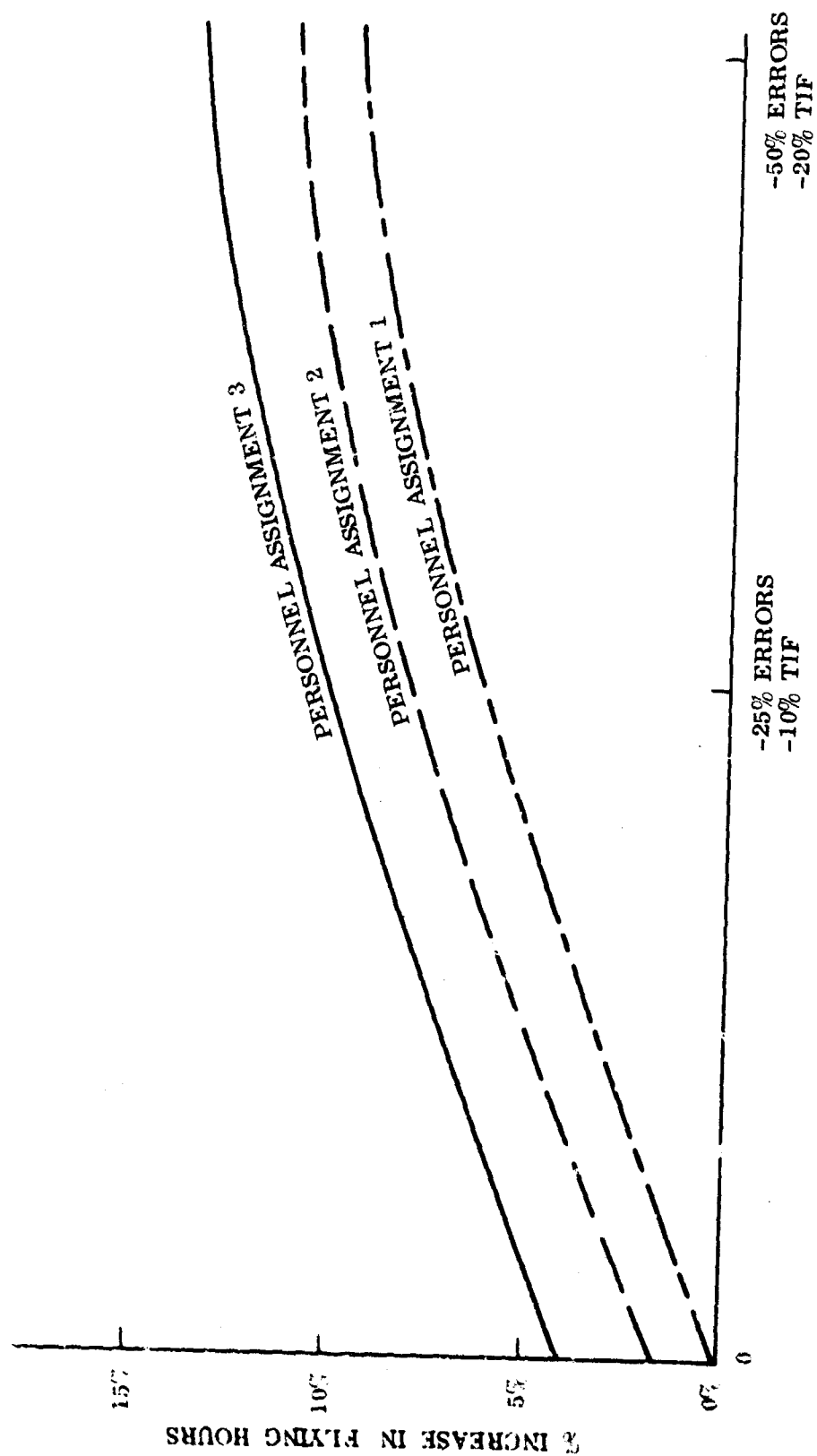POLICY 3 - Lower skill levels used interchangeably with middle level skill levels.

Figure 7-4. Effect of Personnel Assignment on Flight

The justification for these policies is based on earlier field tests, which showed that three-level Air Force technicians using audio-visual information presentation performed as effectively as 5 level technicians using the current T.O. mode. The reduction in time and error rates is also based on field tests results which indicated that a 45% reduction in error rate and a 20% reduction in time could be realized through the use of reformatted technical data presented through the audio-visual mode. As can be seen from this graph, the expected payoff represents an increase of approximately 14% in total aircraft flying hours. At the present utilization rate this means a gain per aircraft month of more than 20 flying hours.

The AMES model was also used to determine the impact of improved performance on the cost of maintenance. One of the measures used in this analysis was maintenance manhours per flight hour. As shown in Figure 7-5, the potential reduction in this variable is approximately 30% using personnel policy 3 and assuming a 50% reduction in errors and a 20% reduction in performance time. Other measures such as spares consumption are used to obtain a total cost saving figure.

In summary, the AMES model has been used as a tool in project PIMO to express changes in maintenance effectiveness resulting from an improved technical data system to changes in effectiveness and cost/effectiveness of the object system, namely the C-141A. The model was constructed to incorporate measures of functional reliability and alternative personnel utilization in a manner consistent with a data collection and field evaluation scheme. The model was used to establish payoffs in terms of increased aircraft utilization and cost savings which could be compared to the cost of information system improvements.

In terms of the PIMO application the model served well as a man/ machine synthesis device, however, this application represents only a subset of the problem areas in which the model could be used effectively. Serendipity is currently pursuing additional areas of application with the Military Airlift Command.
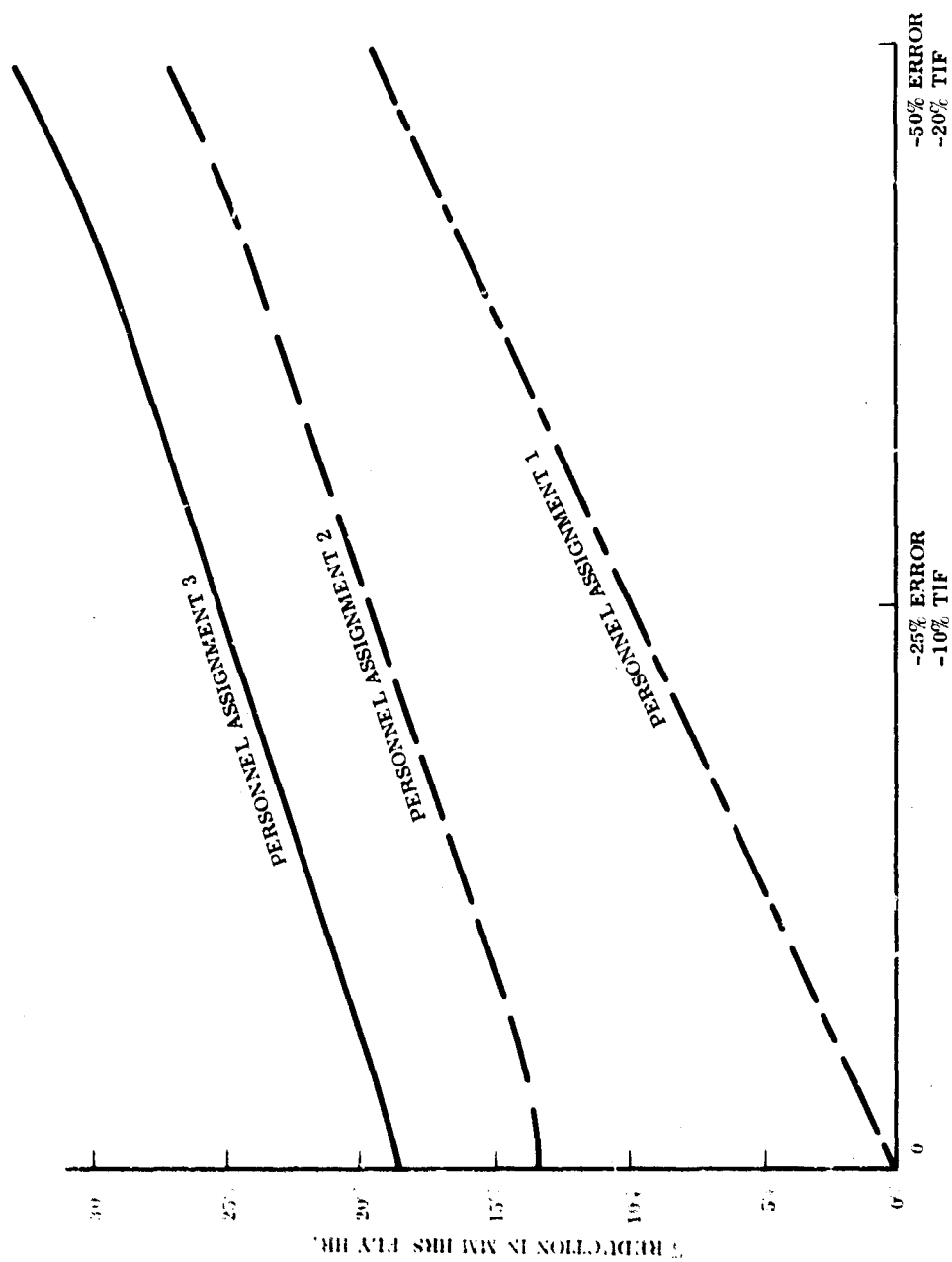
Figure 7-5. Effect of Personnel Assignment on Maintenance

8. MAN-MACHINE SYSTEM EVALUATION—
THE NORMATIVE OPERATIONS REPORTING METHOD

M. Stephen Sheldon and Henry J. Zagorski
System Development Corporation

The rapid development of military man-machine systems in the last decade has presented new problems for people concerned with system measurement and evaluation. Concepts like mean time between failure (MTBF) or circular error probability (CEP) and the classical psychometric approaches are not sufficient to permit adequate assessment of the complex behavior of a system. It is becoming increasingly evident that man-machine system evaluation calls for techniques that are radically different from those which persevere by tradition. We propose that this work area be called systemetrics. As an example of the kind of work that can be done, we are going to describe the Normative Operations Reporting Method (NORM), which is currently being applied in SAGE field evaluation.

The SAGE system represents the first large scale computerized man-machine system in operational use. Our efforts have been intimately associated with the development of SAGE and with various attempts to devise efficient and meaningful methods for the measurement and evaluation of this system. After several years of preparatory work, we have finally succeeded in putting together an evaluation method that works rather well when used in the practical military situation. This paper will describe the method and the context in which it is being applied.

The paper will be organized into four sections. First, the SAGE environment will be described in sufficient detail to allow the reader to gain some appreciation of the measurement and evaluation problems. Then, the SAGE crew performance criterion development procedures will be discussed. The third section will outline in detail the development of the normative evaluation methodology, and in the last section we will try to show the applicability of this methodology in other systems. Before going on, we would like to emphasize that there have been no exotic breakthroughs created during the development of the methodology. The creativity of the method lies in the unique combination and application of assessment techniques that are well known.

An Overview of the SAGE System

The Semi-Automatic Ground Environment, or SAGE, is a computer-based air defense network. It is composed of fourteen direction centers scattered throughout the continental United States. Each of these centers receives raw data pertaining to the air situation in its area of responsibility. These data consist of:

A. Digitalized radar information concerning the up-to-the-minute position of aircraft. This information is transmitted over communication lines to each center from numerous data-linked radar stations.

B. Early warning reports of aircraft tracks transmitted automatically from other direction centers as well as via teletype from airborne or ground early warning stations.

C. Military and commercial flight plans filed with the Federal Aviation Agency and forwarded via teletype to the center.

D.   A variety of intelligence reports, weather reports, airbase and
     weapons status reports and other messages forwarded by automatic
     data-link, teletype or telephone.

At the direction center, these data are all fed into a high-speed digital
computer that processes, integrates and displays relevant operational
information selectively to various members of a specially trained Air Force
crew. The moment-to-moment air traffic situation is displayed via
specialized consoles that are linked to the computer. A wide variety of
displays are available at these consoles. The operators interact with the
computer by means of console switches and light guns in order to direct
the computer to perform certain special routines, such as calculating
desirable intercept tactics against a designated aerial object.

Each direction center is responsible for an air defense area called
a division. The divisions are numerically numbered and are grouped into
regions whose headquarters are called a combat center. Here, another
digital computer receives information that is either forward told from
divisional direction centers or laterally told from adjacent combat centers.
The combat centers process messages concerning the overall air situation
throughout their constituent divisions and in turn forward appropriate
information to the NORAD command control center.

One segment of the operational personnel manning the SAGE direction
center is called the air surveillance section. Here, the operators must
decide which radar data represent actual aircraft and which are due to
noise. When they decide that an aircraft is present, they introduce appro-
priate symbology into the display system by means of console switches and
light guns. This function is called "detection." The air surveillance
personnel are also responsible for a function called "tracking", which
concerns the proximity and appropriateness of the symbology in the display
system in reference to the direction and speed at which the aircraft (radar
returns) are moving in the air space. Although the computer performs most
of the tracking work, the surveillance displays must nevertheless be
monitored continuously for potential discrepancies. When unusual events
such as poor radar data acquisition accompanied by substantial noise occur,
the manual intervention required by the personnel in the air surveillance
section can be considerable.

Once it has been decided that the digitalized radar data represent an
actual aircraft and that the display symbology has been introduced appropri-
ately, this symbology is assessed by personnel whose responsibility it is
to determine the identity of the aircraft. Although there are many con-
siderations and ramifications in the aircraft "identification" function, we
will oversimplify by stating that these personnel essentially decide whether
each aircraft in the display system is Friendly or Hostile. Actually, the
Hostile classification is not used in peacetime operations. Instead, the
identification personnel use the designation "Faker" to classify a make-
believe invader trying to penetrate an air defense area.

When an aircraft has been identified as a Faker, special displays are
directed to personnel in the weapons section of the direction center. These
personnel have two primary functions, first an interceptor must be
committed against the Faker. This function is called "commitment".
Second, the interceptor must be guided appropriately into a position that
will permit the interceptor pilot to take appropriate closing action against
the Faker. This function is called "guidance". There are a variety of
computer routines available to the weapons personnel to assist them in
performing their functions. For example, one routine provides a display
that indicates which interceptor tactic has the best chance of success.

During the guidance process, the computer automatically transmits to the interceptor pilot the successive headings, speeds and altitudes which will give him the maximum probability of making a successful intercept. Needless to say, there are many situations where the personnel responsible for commitment and guidance must intercede in the process in order to achieve successful completion of the weapons functions.

The foregoing description has been a brief and over-simplified account of how the SAGE system operates. The details of the many possible interactions between humans and machines are extremely complex. The computer program alone at each direction center runs into the hundreds of thousands of instructions. This program is in a constant state of maintenance and improvement as the technology of air defense changes.

Criterion Development

Of primary importance in any evaluation methodology is the development of suitable criterion measures. The quality of the criteria will determine more than any single element the meaningfulness of an evaluation. As difficult as it is to achieve valid criteria in a simple situation, it is even more difficult in systemetrics. In dealing with a man-machine system one must ask, "What is the system trying to accomplish?", and, "What available data will adequately reflect system performance?". In air defense, the basic objective of the system is to detect and neutralize invader aircraft before they penetrate designated areas of concern. From the foregoing description of SAGE, we saw that there are five basic functions that are performed in the system: Detection, Tracking, Identification, Commitment, and Guidance. Appropriate decisions and actions associated with these functions must be made by the personnel operating the system. However, the accomplishment of basic functions represents merely one way to look at system performance. For example, an invader aircraft can be detected, tracked, identified, committed against, and an interceptor appropriately guided to the invader AFTER this invader has already penetrated a critical zone. Thus, it is evident that the faster and more accurately the system responds in general, the more effective it is in accomplishing its basic mission.

A precise stipulation of the SAGE system mission which would suggest meaningful performance measures is not to be found. The only generally agreed upon statement of objectives that we were able to formulate can be stated as follows. The system should neutralize as many invaders as possible as quickly as possible and as far out as possible. We translated this overall objective into three quantifiable criterion measures. All measures are calculated at the direction center level.

1.   Percentage Fakers Killed

This measure simply divides the number of Faker aircraft which are neutralized by the total number of such aircraft in a mission.

2.   Faker Life

This measure counts the time that the average Faker is in the division's air space, i.e., from the first time radar is available for it until it is either neutralized or exits from the division's area of responsibility.

3.   Depth of Penetration

This measure averages the depths of penetration of the Fakers into an air defense area.

The above three measures were developed to reflect the basic objectives of the system. However, these measures had to be supplemented by other measures concerning explicit functions performed by the system. Many different measures were examined for possible use at the functional level of performance. Of these, four were able to withstand the testing phase. These are:

1. Detection Latency

    A measure which averages the amount of time between the initial appearance of the Faker and the time the system is made aware of its presence by the initiation of appropriate display symbology.

2. Unassociated Time

    A measure of tracking which averages the time during which the display symbology and the position and direction of the Faker are not in sufficient congruence.

3. Interception Time

    A measure of the time it takes to complete the entire guidance function.

4. Tactical Action Latency

    A measure of the rapidity of commitment function. It represents the average time between detection and the time an interceptor was paired to the Faker.

These measures, along with many others are being collected from simulated air defense missions performed at all SAGE direction centers. The data are obtained from operational recording tapes that contain a history of all relevant activities taking place during a mission. Some card inputs are also used to reduce the data for each mission. A special computer program at each direction center is used to report crew performance and to compile data for ongoing statistical analysis.

In order to develop more comprehensive criteria of effectiveness, the performance measures are being factor-analyzed by the principal components method. To date, the first two factors appear to explain about 76 percent of all the observed variation in performance. The first factor is defined rather well by three measures: Tactical Action Latency, Interception Time, and Depth of Penetration. It seems logical to call this factor Weapons Performance. The second factor is defined by five different measures of air surveillance and is currently called Air Surveillance Performance. These factor scores have been shown to be more reliable than any of the individual measures. They also have intrinsic face validity in that they correspond with the physical organization of the Direction Center.

The criterion research in SAGE has resulted in relevant, quantifiable measures of system and functional performance. The creation of these measures has led to a meaningful procedure for evaluating man-machine performance at the direction center level. This procedure is now built into an operational computer program that is used in the field to assess crew effectiveness immediately after a mission is completed. The program is updated periodically with the aid of appropriate statistical analysis

Evaluation Methodology

It was evident to many observers of the SAGE system that all missions are not of equal difficulty. Different kinds of environments, weapons configurations and invader forces make for missions of decidedly different difficulty. In spite of this, most military evaluation was accomplished by judgment of whether or not certain rigid standards of accomplishment were met. The NORM methodology features a set of flexible standards that are adjusted according to the relative difficulty of the mission.

Once the criterion measures have been defined, the next task in developing the systemetric model is to try to account for that portion of variance in performance that is attributable to the difficulty of the mission. In doing this, it is necessary to determine which characteristics of the mission are most likely to contribute to mission difficulty. In SAGE, the total number of such variables is quite large, however, those which account for significant variance and can be scaled comprise a manageable subset. Over sixty different criterion and predictor variables were investigated by a variety of statistical techniques. Before this could be done, each variable required a very explicit definition that could be translated into a computer program for automatically extracting and formatting the data from the mission recording tapes.

The mission difficulty variables found to be pertinent in SAGE can be grouped into three major classes: 1) radar variables, 2) invader variables, 3) operational environment variables. Explicit examples of these variables are listed below:

1. Frequency of radar returns.

2. Amount of electronic noise.

3. Evasive tactics.

4. Altitude and speed.

5. Nature of air space.

6. Targets being defended.

7. Overall invader load.

8. Relative distances between invaders and interceptor bases.

9. Type of weapons available.

10. Nature of early warning.

Data from these and many other variables were collected, compiled into a computerized data base and analyzed by appropriate statistical procedures. The technical problems involved in formulating, specifying programming, compiling, and analyzing large amounts of field data are not inconsequential. For this reason, the original analysis was performed using individual Faker invaders as data reference points even though averages for missions were much more desirable. When sufficient data became available, mission means for each of the variables being analyzed became the reference points. As anticipated, this change decidedly reduced measurement error variance and increased the precision of evaluation. At present, the data base already contains information for 125 air defense missions representing all the SAGE direction centers in the system. All formal

statistical analysis is done via FORTRAN routines using the 7094 computer. In addition, computer time-sharing statistical procedures are being investigated for feasibility of application to this project.

The overall purpose of the systemetric model is to develop a methodology that permits an evaluation of man-machine performance based upon a series of flexible standards reflecting the difficulty of the mission. This approach is in direct contradistinction to the absolute standards approach. In order to develop such a set of standards, it is necessary to be able to estimate with reasonable accuracy how well an average crew will do on any measure when performing a mission of known difficulty. In other words, it is necessary to be able to predict the performance of a crew on the basis of how hard the mission is. If an "expected score" on each criterion measure can be developed for an average crew based on relevant mission difficulty variables, this score can be compared to an "observed score" and the residual, or difference, can be used as a basis for evaluation. This is the kind of evaluation being accomplished for SAGE by the Normative Operations Reporting method.

Initial statistical procedures in NORM focus on the basic correlations between each criterion measure and each potential mission difficulty variable and on the relative independence of the variables being considered as predictors. This is followed by a series of multiple regression runs for each measure using selected sets of mission difficulty variables as independent variables. The final selection and weighting of these variables is made on the basis of exhaustive analysis; including such considerations as quality of distribution function, statistical validity, independence, face validity, reliability and accessibility of data, and the reasonability of assuming that a variable does indeed account for performance variation. Overall, about 50 percent of the variance of criterion performance is being accounted for by the presently available mission difficulty variables. Table 8-1 gives the multiple R, standard error and percent variation accounted for in each criterion measure now being used.

TABLE 8-1
STATISTICAL SUMMARY (Based on 93 Missions)

Measure	Multiple Correlation Coefficient	Percent Variation Accounted For	Standard Error
Percentage Fakers Killed (%)	0.58	34%	10.38
Faker Target Life (min.)	0.81	66%	5.93
Weapons Performance (factor score)	0.81	66%	6.13
Tactical Action Latency (min.)	0.75	56%	1.24
Interception Time (min.)	0.77	60%	4.32
Depth of Penetration (n. m.)	0.89	80%	33.39
Air Surveillance Performance (factor score)	0.67	45%	7.68
Detection Latency (min.)	0.58	34%	2.05
Unassociated Time (min.)	0.63	40%	2.38

The procedures used to select predictors and accomplish multiple regression analysis would probably offend the statistical purist. For example, variables, with very low or even reverse sign validities are sometimes included in the prediction equations because their beta weights are in the appropriate direction and they possess a strong intuitive relationship to performance. This procedure was used to select electronic noise as a predictor in evaluating the detection and tracking functions. It was intuitively obvious to the research staff that the more noise is present in a display the more difficult it is to detect and track the actual aircraft. In spite of this observation, the basic correlations between the noise variable and the detection and tracking measures were generally low and in the wrong direction. Since the beta weights turned out properly, it was inferred that the basic correlations were affected adversely by the confounding of difficulty variables in field operations. Another consideration in the inclusion of the noise variable in the prediction equations was to hedge or guard the evaluation model against situations where considerable noise is being introduced to train and test crews. A number of other variables of this type are included in the prediction model to protect it against extreme conditions. In addition, there are numerous other devices used to prevent these equations from assuming unreasonable values. Legal limits are defined and set for each predictor and prediction. The technique called Winsorization is used generally throughout the model to control the prediction of expected performance.

The systemetrics approach requires the researchers to be intimately familiar with the system, to know the meaning and importance of variables as well as their statistical characteristics and to have a knack for selecting and using variables in ingenious ways to meet the objectives of system evaluation.

Validation of Normative Evaluation Methodology

Having achieved initial success in predicting and evaluating performance, an experimental computer program is now being used in all SAGE direction centers to further validate the methodology. This program is run at the conclusion of each air defense mission. It reads the mission recording tapes and outputs an expected score and an observed score for each criterion measure. Then, it determines the difference between these scores, divides the differences by appropriate error terms, converts the resulting ratios into performance stanines, and produces an appropriate evaluation of performance for each criterion measure as well as total performance. One page of output contains the name of each measure, the observed and expected scores and the stanine presented numerically, graphically by a bar diagram and verbally by phrases ranging from "very good" for a 9 to "very poor" for a 1. A facsimile of this computer-generated performance report appears in Table 8-2.

A second output page lists all performance measures and mission difficulty variables being used or under consideration along with their mission mean values. The program user, who is normally the training officer, is required to make a subjective input to the program concerning the reputed skill of the crew. This rating is printed on the second output page along with other identifying information. The crew skill rating is trichotomized into 1) Highly Skilled, 2) Average and 3) Trainee. The missions manned by average crews are used as additional data points to develop the equations. The missions manned by highly skilled crews and trainee crews are used to further validate the evaluation model.

Although subjective corroboration of field evaluations by on-site observers is being used to some degree to further validate the evaluation

methodology, the primary test being used at the present time is the extent to which the methodology discriminates between expert crews and trainee crews. So far, on the basis of limited results, the method appears to discriminate such crews rather well. The average stanine difference between these types of crews is 2.5 of which about 2/3 is contributed to by sheer differences in raw criterion scores and 1/3 by differences in mission difficulty. It is anticipated that with additional data and more accurate methods for ascertaining the overall skill rating of a crew, these results should become even more conclusive. At this time, the incremental accuracy and efficiency of evaluation afforded by the Normative Operations Reporting Method appears to be significant.

Applications of Normative Evaluation Methodology

This paper has been concerned with an approach to the measurement and evaluation of systems called systemetrics. More specifically a system evaluation method developed out of this approach, called the Normative Operations Reporting Method (NORM), has been described as it is being applied to the SAGE system of air defense. Because the method has been demonstrated to have adequate validity and to be acceptable to military users, it is believed to have a potentiality for application in other operational situations.

An example of a potential hardware application is in the radar systems area. Here, the standard criteria for evaluation have been range and azimuth accuracy. Seldom does the evaluation consider the electronic environment in which the radar system is being or will be used. Furthermore, the evaluation does not take into account variables describing human factors, weather, logistics, altitude, antenna position, and a host of other conditions which can potentially affect how well the system will perform. The criterion development procedures and normative evaluation methodology described here appear to have ready transferability to the evaluation of radar and other hardware systems.

Another area which badly needs systemetric development concerns the various educational systems. Teachers, schools, and school districts have characteristically avoided comparative evaluation, claiming that each school situation is unique and consequentially incomparable with any other. An appropriately conceived normative evaluation model should be able to overcome these objections and make comparative assessment possible. There is no doubt that suitable criterion measures can be developed for educational evaluation.

These criterion measures should then be normatively calibrated to take into account such things as pupil/teacher ratio, operating cost per pupil, teacher salaries, and numerous other potentially relevant variables.

Other military systems, communication systems, and industrial systems appear to be ready markets for the systemetric approach. With the high speed computer as the supporting tool, the number of variables that can be considered in statistical analysis is no longer a real constraint to the energetic scientist. The measurement and evaluation of systems by means of systemetrics can and should become an important part of the work of the human factors scientist.

TABLE 8-2. Norm Performance Report

Division XX   Mission XXXXXX   Date XXXXXX

Measure	Observed Performance	Expected Performance	Stanine	Profile 1 2 3 4 5 6 7 8 9	Evaluation
Percentage Fakers Killed (%)	92	93	5	X X X X X	Average
Faker Target Life (min.)	21	27	7	X X X X X X X	Above Average
Weapons Performance (factory score)	105	101	6	X X X X X X	Average-High
Tactical Action Latency (min.)	4	3	3	X X X	Below Average
Interception Time (min.)	15	14	5	X X X X X	Average
Depth of Penetration (n.m.)	89	122	7	X X X X X X X	Above Average
Air Surveillance Performance (factory score)	110	98	8	X X X X X X X X	Good
Detection Latency (min.)	2	3	6	X X X X X X	Average-High
Unassociated Time (min.)	3	8	9	X X X X X X X X X	Very Good

OVERALL PERFORMANCE PERCENTILE 65